Community Expectations for Research Artifacts and Evaluation Processes

Ben Hermann
ben.hermann@upb.de
Heinz Nixdorf Institut
Universität Paderborn
Paderborn, Germany

Stefan Winter
sw@cs.tu-darmstadt.de
Dependable Systems and Software
Technische Universität Darmstadt
Darmstadt, Germany

Janet Siegmund
janet.siegmund@informatik.tu-chemnitz.de
Technische Universität Chemnitz
Chemnitz, Germany

ABSTRACT

Artifact evaluation has been introduced into the software engineering and programming languages research community with a pilot at ESEC/FSE 2011 and has since then enjoyed a healthy adoption throughout the conference landscape. In this qualitative study, we examine the expectations of the community toward research artifacts and their evaluation processes. We conducted a survey including all members of artifact evaluation committees of major conferences in the software engineering and programming language field since the first pilot and compared the answers to expectations set by calls for artifacts and reviewing guidelines. While we find that some expectations exceed the ones expressed in calls and reviewing guidelines, there is no consensus on quality thresholds for artifacts in general. We observe very specific quality expectations for specific artifact types for review and later usage, but also a lack of their communication in calls. We also find problematic inconsistencies in the terminology used to express artifact evaluation's most important purpose – *replicability*. We derive several actionable suggestions which can help to mature artifact evaluation in the inspected community and also to aid its introduction into other communities in computer science.

CCS CONCEPTS

• **General and reference**; • **Software and its engineering** → *Software libraries and repositories*; *Software verification and validation*;

KEYWORDS

Artifact Evaluation, Replicability, Reproducibility, Study

ACM Reference Format:
Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community Expectations for Research Artifacts and Evaluation Processes. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20), November 8–13, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3368089.3409767

1 INTRODUCTION

In 2016, a replicability crisis became public, when more than 1500 researchers revealed having trouble replicating previous research results [1]. This replicability crisis also reached the software engineering community, as it has embraced the importance of replication for knowledge building [3, 4, 15, 21, 22]. For example, Collberg and Proebsting could not obtain the relevant artifacts to conduct a replication, neither by contacting the authors, the authors' institution, and funding agency [7]. Also, Lung et al. describe their difficulties in conducting an exact replication, even when they were in direct contact with the authors [17]. Glanz et al. describe similar experiences when obtaining research artifacts for comparison and had to reimplement competing approaches in order to replicate results [10]. For the term *artifact*, we follow the definition provided by Méndez et al. [18], describing it as a self-contained work result with a context-specific purpose.

To improve the situation of missing or unusable artifacts, artifact evaluation has become a regular process for scientific conferences in the software engineering and programming language communities. It contributes to the larger trend towards open science in computer science. Since the first piloting of the process at ESEC/FSE 2011, many other conferences have included artifact evaluations as an additional step that authors of accepted papers may take. If their artifact is successfully evaluated the corresponding publication is marked with a *badge* [9, 11] indicating different levels by which the artifact is found to support the presented research results. Successfully evaluated artifacts are listed on the conference website and commonly linked with the paper in publication repositories such as the ACM Digital Library. Except for few venues (i.e., CAV and TACAS), where artifact evaluation is mandatory for tool papers, artifact submission usually is a voluntary activity that authors of accepted publications are invited to participate in. Journals are recently adopting the idea of artifacts as part of open science initiatives. For example, the Empirical Software Engineering journal (EMSE) encourages authors to share their data in a replication package [19]. There is preliminary evidence that papers with an evaluated artifact have higher visibility in the research community [6, 13].

There is, to the best of our knowledge, currently no evidence that artifact evaluation is leading to better artifacts for computer science research communities. The overarching goal of our work is to enable an assessment of the efficacy of artifact evaluations as they have been implemented in software engineering and programming language conferences and to identify possible improvements for these processes. Such an assessment requires criteria according to which we can judge whether artifact evaluations meet their

ESEC/FSE '20, November 8–13, 2020, Virtual Event, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7043-1/20/11…$15.00
https://doi.org/10.1145/3368089.3409767

objectives. However, from an initial review of the ACM's guidelines on artifact review and badging [9] and the different conferences' calls for artifacts we were not able to derive clear and uniform criteria what makes a research artifact "good". The standard of quality widely varies between different conferences and evolves over time. Thus, the quality of artifacts of different venues is not necessarily comparable, making it difficult to reach a unified quality standard that artifacts should adhere to.

As a first step towards a systematic assessment of artifact evaluation processes, the objective of this paper is to assess their current perception in the AE-pioneering software engineering and programming language communities and to pave the way to unified quality standards regarding artifact evaluation. To this end, we qualitatively examine **(RQ1)** *the purpose of artifact evaluation*, **(RQ2)** *the quality criteria and expectations for research artifacts*, and **(RQ3)** *the magnitude of difference in the perception of purpose and expectations within the software engineering and programming languages communities*

To answer these questions, we have conducted a survey among researchers who have served on artifact evaluation committees (AECs), as they have experience with the expectations toward artifacts and the procedural challenges. We have contacted all members of AECs, including the respective chairs, for all artifact evaluations conducted at software engineering and programming language conferences between 2011 and 2019.

We found that the perceived purpose of artifact evaluation is to foster replicability and reusability at the same time. While we could observe several quality criteria to be expected from artifacts, we found no clear consensus on them. Moreover, the expressed expectations of the communities are largely not represented in the calls for artifacts. This makes it hard to define a quality standard for an individual conference, the community, or a cross-community quality standard. The results of our study show that the lack of such quality standards leaves reviewers without guidance how to decide on artifact acceptance or rejection. Moreover, it creates an ambiguity for readers of research articles how to interpret the badges awarded to papers after AE.

From these observations we derive the suggestion that commitees should be instated in the programming language and software engineering communities to drive and foster the clarification of the purpose of artifact evaluation within the respective community, along with corresponding review guidelines.

In summary, we make the following contributions:

- We provide an overview of the current perception and practice of artifact evaluation and the expectations toward artifacts and the process.
- Based on community inputs, we present suggestions for future development and improvement of artifact evaluations.
- We published the survey, data set, scripts, and analysis results that our conclusions are based on as a research artifact for replicability of our results, for further analysis, and for extension by the community [12].

2 BACKGROUND AND RELATED WORK

The concept of artifact evaluation as a means to foster replicability is a relatively new practice in software engineering research. It has also been discussed in a broader computer-science community in a Dagstuhl Perspectives Workshop (15452) in 2015, where one of the key results was that the community needs to be pushed further to embrace the publication and—most importantly—sufficient documentation of artifacts.

Méndez et al. found that there is no agreed-upon understanding of what an artifact actually is [18], so they set out to explore potential definitions. They come to a general definition that we also adhere to in our work: "An artefact is a self-contained work result, having a context-specific purpose and constituting a physical representation, a syntactic structure and a semantic content, forming three levels of perception".

Replication in software engineering has become more and more important. Already 20 years ago, Basili et al. found that "too many studies tend to be isolated and are not replicated, either by the same researchers or by others" [3]. To support replication they developed a framework for describing related studies to allow researchers viewing them in context rather than in isolation. Despite the difficulties of actually conducting replications, as reported by Lung et al. [17], Shull et al., as well as Juristo et al., have pointed out the importance of replications [15, 21]. Both encourage the software-engineering research community to embrace replications because the context of human studies in software engineering is too complex to be understood with a single study. However, as Siegmund and others point out, the community has to overcome the hypocrisy of paying lip service to the importance of replication but at the same time not valuing them accordingly [22].

Robles reported very scarce availability of artifacts in the Mining Software Repositories (MSR) community between 2004 and 2009 impeding replication of results [20]. As this community within the software engineering field was primarily focussing on the use and reuse of datasets it was reliant on the availability of datasets. While artifact evaluation was piloted later in 2011, in 2005 Tim Menzies and Jelber Sayyad started the now discontinued PROMISE repository[1] to share research artifacts. Artifacts were archived without a formal review process. They received the MSR Foundational Contribution Award in 2017 for their work.

Wacharamanotham et al. inspected the low availability of artifacts in the HCI community and found that four factors influence researchers to refrain from sharing artifacts: concern about personally-identifiable data, lack of participant's permission, lack of motivation, resources, or recognition, and doubt in the usefulness of their artifact outside their own study [24]. Dahlgren conducted an observatory study during the OOPSLA 2019 artifact evaluation and found that the most prominent negative comments during artifact review are due to limited physical resources or review time to test artifacts and problems with documentation [8].

Timperley et al. conducted a survey among authors of published papers at ICSE, ASE, FSE, and EMSE 2018. [23]. Together with a publication analysis they studied the current practice and the problems involved in artifact sharing in the software engineering community. In their results, they report similar findings as Wacharamanotham et al. found for the HCI community which suggests that artifact sharing has comparable issues throughout computer science. They

[1] The artifacts have been moved to Zenodo for long time archiving. https://zenodo.org/communities/seacraft

derive several recommendations for different stakeholders in research which align with the recommendations we make in this paper.

3 EXPECTATIONS IN THE ACM GUIDELINES AND CALLS FOR ARTIFACTS

In a pre-study we analyzed the ACM guidelines for artifact review and badging [9] and calls for artifacts (CfAs) issued for software engineering and programming language conferences between 2011 and 2019[2].

3.1 Methodology

To extract expectations on artifacts from these text sources, we analyzed the texts for explicit statements of two types: (1) Statements about the *purpose* of artifact evaluations as a process and (2) statements about *criteria* that artifacts under evaluation are expected to meet. The analysis was performed manually by one researcher and confirmed by another one independently. A tool for plagiarism checking[3] and a tool for difference visualization[4] were used to aid the analysis in order to recognize repeating passages. We expect the stated criteria to follow from the stated purpose, however, analyzing both kinds of statements allows us to identify possible inconsistencies. Such inconsistencies would indicate possible misunderstandings of the used terms, be it on our side or on the side of the calls' authors.

3.2 Results

3.2.1 Expectations on Artifacts in the ACM Guidelines. While the ACM guidelines do not make an explicit statement regarding the purpose of artifact evaluations, they motivate it by an observed lack of reproducibility of research results and define three different desirable properties of experimental research results: *repeatability* (same results if repeated by the same team with the same setup), *replicability* (same results if repeated by a different team with the same setup), and *reproducibility* (same results if repeated by a different team with a different setup)[5]. Repeatability is stated as a minimum requirement for publishing experimental results, reproducibility as "the ultimate goal", and replicability as the intermediate property targeted by artifact evaluations. Krishnamurthi and Vitek [16] name *repeatability* as the primary goal of artifact evaluation and describe it as re-running a bundled software artifact. This is in essence what the ACM guidelines now describe as *replicability*. We believe this to be a case of terminology evolution as Krishnamurthi and Vitek do not make the explicit distinction of the group performing the experiment repetition the more recent ACM guidelines make.

The ACM guidelines state criteria that artifacts must fulfill in order to be awarded one of five different badges. Three badges are recommended to be issued in the context of artifact evaluations: "Artifacts Evaluated – Functional", "Artifacts Evaluated – Reusable", and "Artifacts Available". The first two badges require the artifact to have passed an independent audit with different criteria. For the functional badge, an artifact needs to be "documented" (sufficient description to be exercised), "consistent" (contributes to how paper results were obtained), "complete" (includes all components relevant to the paper to the degree possible), "exercisable" (executability of scripts/software, accessibility/modifiability of data), and "include appropriate evidence of verification and validation". For the reusable badge, the artifact must meet the functional badge's criteria and, in addition, must be particularly well documented and well structured to facilitate *reuse* and *repurposing*. For the available badge, artifacts need to be made publicly accessible on archival platforms with "a declared plan to enable permanent accessibility". Two other badges are proposed for papers, for which the main results have been replicated or reproduced in subsequent studies according to the definitions set forth by the guidelines.

3.2.2 Calls for Artifacts (CfAs). Contrary to the ACM guidelines, 61 out of 79 analyzed calls for artifacts explicitly state a purpose for artifact evaluation. Across all analyzed calls, the most frequently named purpose of artifact evaluation processes is to enable *reuse* of artifacts (32 calls[6]), followed by *reproducibility* (24) and enabling *comparison* (17) for future research against published results. When divided by community, programming languages conferences[7] named reproducibility (21) more often than reuse. Some of the calls name the weaker properties of *replicability* (6), which is the declared direct goal of artifact evaluations in the ACM guidelines, and repeatability (4). Other calls contained more vague statements regarding the purpose (e.g., to provide "evidence for quality" or "support for the paper". Seven calls attribute benefits in terms of reproducibility, replicability, or repeatability explicitly to the *availability* of artifacts, in which they see a purpose of artifact evaluations.

While analyzing calls we noticed suble differences in the use of the terms *replicability* and *reproducibility*. As discussed previously, we believe this to be partly an issue of terminology evolution. However, the notions are discussed inconsistently in the literature as well. While the ACM guidelines refer to a definition from the "International Vocabulary of Metrology" [2], another widely referenced definition is found in the ASA's "Recommendations to Funding Agencies for Supporting Reproducible Research" [5], according to which research is reproducible if performing identical data analyses on identical data yields the same findings. Result replication, according to the ASA definition, requires the repetition of a study independent from the original investigators and without using the original data. Although the ACM guidelines on AE provide clear definitions for the terms to be used in the context of AE, both definitions that assign reciprocal meanings to reproducibility and replicability are widely used. We recommend that AEC chairs make this distinction explicit in CfAs to avoid misunderstandings in the interpretation of CfAs and in discussions among AEC members.

Concerning the artifact criteria stated in the calls, we distinguished between evaluation criteria and submission criteria. While evaluation criteria describe properties of the artifact itself, submission criteria are concerned with formal requirements of additional

[2] The corpus of CfAs can be found on our web site https://bhermann.github.io/artifact-survey/ and in our artifact [12].
[3] https://github.com/diogocabral/sherlock
[4] `git diff -no-index -color-words`
[5] The ACM guidelines have been changed after our article has been accepted for publication and now assign reciprocal meanings to the replicability and reproducibility terms (and related badges). We have chosen to not alter the discussion in the paper, as the meanings that were *originally* assigned to these terms were what we expected to be reflected in CfAs and our survey participants' replies.
[6] Numbers do not sum up to 79. Multiple purposes may have been named by one call.
[7] OOPSLA, PLDI, POPL, ECOOP, SAS, SLE, PPoPP, CGO, ICFP, TACAS

material required to submit the artifact for evaluation. An astonishing number of 14 calls does not state explicit evaluation criteria for artifacts, spanning conference calls from 2011 until 2019. We assume that detailed evaluation guidelines were communicated by other means to AEC members. The most prevalent criteria are for the largest part paraphrased from the ACM guidelines (or copied from calls that later heavily influenced the ACM guidelines): *documentation* (46), *consistency* (45), *completeness* (39), and *reusability* (36). Eight calls even contain verbatim copies of the corresponding criteria definitions from the ACM guidelines. On the one hand, this is a clear indication that this set of criteria has evolved as a community standard which serves as a framework for artifact evaluations. At the same time, these criteria do not define clear conditions or thresholds to decide for artifact acceptance or rejection, as the ACM guidelines acknowledge.

In terms of submission criteria in the calls, we find that 22 calls from conferences between 2012 and 2019 do not state any explicit submission criteria. In the calls that do, the most frequently stated criteria are all related to documentation: How to replicate paper results (21), how to use the artifact (17), and how to conduct setup and basic testing of the artifact within less than 30 min (10). This is remarkable for two reasons. First, while the most frequently stated *purpose* of artifact evaluation is reuse, the most frequently stated *submission criterion* is documentation for replication. Together with the observation that consistency is more frequently stated as an evaluation criterion than reusability, this may indicate that the assessment of replicability actually plays a more important role in the artifact evaluation process than the assessment of reusability. Second, the time limit for setup and basic tests is the only statement of an actual threshold we find across all criteria stated in the calls.

In our analysis of the calls we noticed that a good fraction (59) of the calls exemplary name diverse *types* of research artifacts that may be submitted to the AE track, e.g., *code and software* (49), *data* (43), *proofs* (25), but also *grammars*, *surveys*, and even *hardware*. On the one hand, most calls (34) explicitly state that these lists of types are not exhaustive and to be understood as examples. On the other hand, listing these types of artifacts (vs. others) indicates certain expectations of the AEC chairs what they will be evaluating in the AE process. Interestingly, we found only two calls (CAV 2018, VISSOFT 2019) that explicitly state evaluation criteria for (some) types of artifacts they list. 39 calls state specific submission criteria for artifacts of certain types (mostly for code in SE CfAs and for data in PL CfAs). However, these criteria only cover formats to be used, e.g., csv/json/xml for data artifacts, tar/zip for source code, or Docker/VMs for executable software. Therefore, while calls often distinguish different artifacts types, they do not make distinctions in the criteria that apply for these artifact types.

In summary, it is unclear from the calls, (a) from an artifact submitter perspective how to best prepare an artifact so that it is positively evaluated and (b) from a potential (re)user of an artifact what to expect from an evaluated artifact. However, during the artifact evaluation process, criteria to decide on acceptance or rejection *must* have been used. Therefore, we decided to conduct a survey among the AEC members who made these decisions to obtain a better understanding of these criteria.

4 EXPECTATIONS OF ARTIFACT EVALUATION COMMITTEE MEMBERS

To investigate which of the expectations toward artifacts that we extracted from the calls are considered of particular importance and to capture expectations beyond what is expressed in the calls, we conducted a survey across AEC members.

4.1 Methodology

Objective. Based on our results from the analysis of the ACM guidelines and CfAs, we designed our survey to cover four aspects: (1) The purpose of AE (**RQ1**), (2) expectations toward artifacts as a reviewer on an AEC (**RQ2**), (3) expectations toward artifacts as a user after successful evaluation (**RQ2**), and (4) other quality factors of artifacts the participants have (**RQ2**).

Survey Questionnaire. In addition to these core aspects, we also asked the participants about their experience with artifact evaluation and how useful they find the ACM policy to guide their evaluation of artifacts. This helps us to understand the experience of participants with artifact evaluation and to set their responses to later questions into perspective. Also, to answer **RQ3** we asked the participants to specify the AECs they served on. The questions were organized in two main groups separating questions relating to artifact evaluation from those relating to artifact usage. Questions were stated deliberately open so participants could freely share their views. Questions with numerical answers were accompanied by a text field for further elaboration. The full questionnaire can be found on our project web site[8] and in our artifact [12].

Survey Pre-Test. We piloted and refined our survey in several steps with 6 participants, ensuring that we ask the right questions with an unambiguous wording.

Participants. We sent the survey to 1034 members of artifact evaluation committees of different venues and different years. Figure 2 shows a histogram of individuals by the number of AECs they served on. We only included committee members of already completed artifact evaluations at the time of our survey. For FSE 2012, we found a call for artifacts, but could not find a public list of committee members or chairs. For FSE 2013, we could only identify the chairs, whom we also included.

Participants needed a median of 21 min and 19 s to complete the survey. All in all, 257 committee members responded, of whom 124 completed the entire survey. 133 did not complete the entire survey, but we still included their relevant perspectives to answer RQ2 and RQ3. We have excluded the complete replies from two participants from our analysis; one for obviously implausible replies (ID 99 in our data set) and one for the fact that the participant indicated in answers not to feel qualified to answer the questions (ID 218). Out of the remaining 255 responses, 152 indicated the AEC that the respondents served on. We classified the conferences as belonging "more" to the software engineering (SE) or programming language (PL) community as stated in 3.1 and used the 152 (125 for PL, 36 for SE) responses to answer RQ3. Nine respondents had indicated having served on both PL and SE AECs and we include their responses in the analyses for either community.

[8] https://bhermann.github.io/artifact-survey/questionnaire/

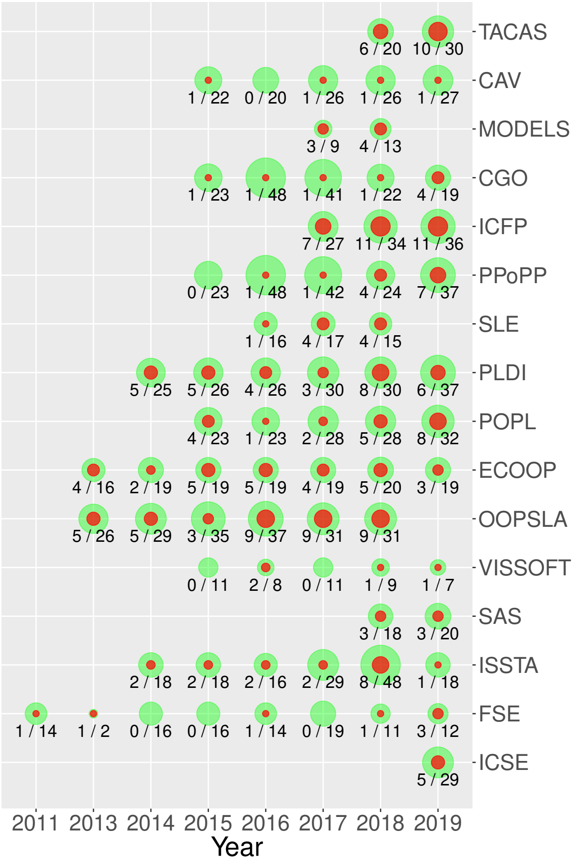

Figure 1: Committee sizes (green) and responses (red) by conference and year

Analysis Protocol. We followed Hudson's approach of open card sorting to analyze the answers [14]. We assigned (at least) two authors to process each survey question. One author identified higher-order topics to each answer. As the process was open, there were no predetermined categories, but they were extracted while reading the answers. For instance, for the answer "Reproducibility to a certain exten[t]. Availability of the code." to the question "[...] what is the purpose of artifact evaluation?" the labels "reproducibility" and "availability" were extracted. The other author checked the labels. Difficult cases were marked and discussed with all authors until consensus was reached. In a second pass, we reviewed all assigned labels and simplified/harmonized labeling, as different authors had used different labels for the same concept.

In the following, we will also present verbatim quotes from respondents. For better contextualization we indicate the respondent ID and their frequency of AEC membership separated by communities if provided by the respondent.

4.2 Perceived Purpose of Artifact Evaluations

To address our first research question, we asked our participants to describe their view on the purpose of artifact evaluation. We received 147 answers.

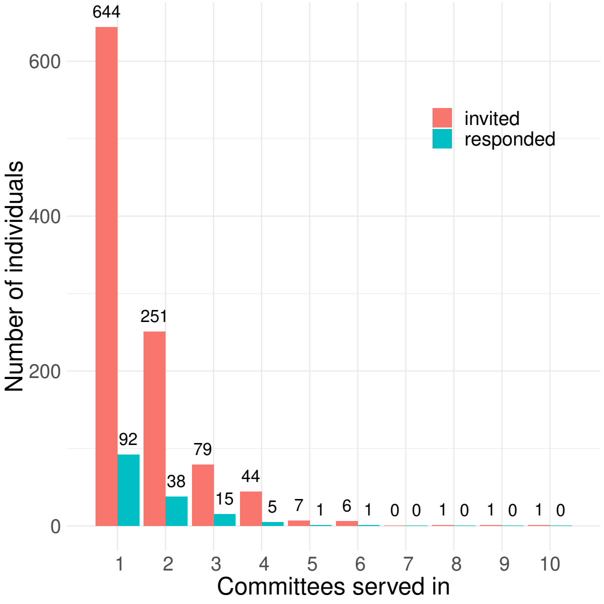

Figure 2: Histogram of individuals by number of AECs served in

4.2.1 Results. In the mentioned purposes, two major groups occurred: *Fostering certain properties of the artifact* and *Checking certain properties of the artifact.* We describe the properties to be fostered or checked in the following.

Fostering Properties of Artifacts. In the first group of answers regarding the fostering of certain artifact properties, the following properties were mentioned frequently: *Reproducibility* (34), *reusability* (26), *comparability* (5), *repeatability* (5), *replicability* (5), *usability* (4), and *availability* (4).

However, as we found in Section 3.2 that *reproducibility* has an inconsistent interpretation across the different calls, we assume most participants in our study actually mean *replicability.*

> In the context of reproducible science contributions, it is important to promote artifacts of scientific quality. That said, artifact evaluation has the goal of validating the quality of artifact in order to guarantee various properties that increases the chances of reproducibility of the experiment over time (eg months, years, centuries...).
> Another aspect of artifact evaluation is, in my humble opinion, the promotion of artifact as first-class scientific contribution, with a recognition by peers as complementary, if not equivalent, in quality and value, to published papers.
> (id 220, 1 SE AEC, 2 PL AECs)

Next, the second most frequent opinion is that artifact evaluation *fosters reusability.* Reusability in this context means that researchers will be able to reuse an artifact of a different research group possibly in a slightly different context or to build upon it for further research. One of our participants summarizes this dual purpose as follows:

> I see two main objectives of artifact evaluation: (1) tempering the tendency to over-promise and under-deliver and (2) incentivizing the ability to build on others' research. [...] (id 48, 3 PL AECs)

Checking Properties of the Artifact. In the second group of properties concerning checks/validations of an artifact the following were frequently mentioned: *Validating claims* (28), *validating results* (23), *validating reusability* (9) *validating reproducibility* (9), *validating existence* (6), *validating replicability* (3), and *validating usability* (3).

> The purpose is to assess that the submitted paper is supported by actual tools and experiments, and that these experiments can be run again in a self-contained environment to reproduce the paper's results. A more ambitious goal is to provide an environment in which the provided experiments can be modified easily (e.g. modifying a test case, commenting parts of a benchmark file, etc.) to see how the tools handle such changes, and how robust the experimental results are. (id 268, 2 SE AECs)

The most mentioned objective for artifact evaluation is the *validation of claims* made in the paper or its *results*. Interestingly, both objectives *validating reproducibility* and *validating reusability* do not seem to be important for participants, even though these objectives are the primary properties artifact evaluation should foster.

4.2.2 Discussion. Similar to what is stated in CfAs, the participants see the mission of artifact evaluation in fostering replicability and reusability. Contrary to what we observed for CfAs, replicability (or "reproducibility") is mentioned by a larger number of respondents than reusability, which likely is an effect of the sample of AEC members that we received responses from. The majority of respondents stated to have served on PL AECs (131 vs. 43 on SE AECs with 12 respondents having served on AECs for either community), for which replicability is the most frequently stated AE purpose in CfAs. Also, the role of non-exact replications [15] has reached the research community, as participants mentioned that they would like to be able to alter the experimental setups provided with the submitted artifacts, so that they can test their robustness.

However, when speaking about the validation of properties of artifacts, the most frequently mentioned property is the *validation of claims or results*. This is still very close to the original mission to hold the artifact accountable to the expectations set by the paper, which is what Krishnamurthi and Vitek report [16].

> While the direct purpose of artifact evaluation with regard to the submitted artifacts is the validation of results or claims made by the paper, the community has extended this initial mission of artifact evaluation and now also sees its purpose in *fostering* replicability and reusability.

We also find that terminology is not used consistently among participants, similar to our finding for inconsistent terminology in calls (subsection 3.2). Specifically, many participants wrote about reproducibility when they actually meant replicability. However, to clearly communicate the expectations toward artifacts, we need to decide on a consistent terminology.

> Terminology for the most important purpose of artifact evaluation is used inconsistently in the community.

4.3 Expectations of the Community

From our analysis of the replies we received regarding RQ1 (see Section 4.2), we have seen that the purpose of artifact evaluation is perceived as two-fold: Verifying the accuracy of claims and results in research articles and (re-)usability of artifacts. In our analysis of CfAs, we found an indication that reusability may play a lesser role in terms of *evaluation criteria* for artifacts. To further investigate this hypothetical finding, we collected two distinct perspectives from our participants, first as a reviewer, and second as a user. If the expectations differ for these two perspectives, this would support the result from our call analysis.

As the evaluation criteria in CfAs were rather unspecific for the largest part, we also asked specific questions regarding evaluation criteria for the types of artifacts that are most frequently named in calls (code/software, data, proofs) to obtain a better understanding of the actual decision criteria for artifact acceptance and rejection (RQ2).

In our analysis we differentiate between respondents that have served on PL and SE AECs to address RQ3. Please note that the total numbers reported can be higher or lower than the sum of PL and SE responses, because (a) not all respondents provided information on which AECs they served and (b) respondents may have served on AECs in both communities.

4.3.1 Perspective as Reviewer.

Expectations in General. To understand the quality criteria that *reviewers* expect from an artifact, we asked for the minimum requirements to accept an artifact (124 answers) and for the reasons to recommend an artifact for acceptance or rejection (110 answers).

The most frequently mentioned criteria were *replicability of results* (45) (PL/SE: 39/7), *good documentation* (43: 32/11), and *easy setup* (37: 26/9). Several participants mentioned that they accept artifacts that show some *"general" replicability* (12: 8/2). Looking at responses from the SE community in isolation, replicability is only ranked third; good documentation and an easy setup are perceived to be more important. We suspect this to be an indication that the SE community values reuse over replication. However, we could neither confirm nor refute this based on our data.

18 (15/2) participants reported that they recommended accepting an artifact because they *were able to replicate results*. 14 (10/4) participants reported that they recommended rejecting because they *were not able to replicate results*. Further reasons for acceptance suggestions were: *easy setup* (7: 4/1), *good documentation* (5: 4/–), *matches with the claims from the paper* (5: 5/–), *meets minimum requirements* (5: 5/–). Further reasons for rejection were: *bad documentation* (5: 4/1), *results deviate too much from the ones reported in the paper* (5: 4/–), the artifact was *substantially different from the paper* (4: 3/1). We discuss the most frequently mentioned criteria in more detail in the following for a more detailed description what respondents mean by these terms.

Replicability of Results. As in the responses for the purpose of artifact evaluation, we saw an inconsistent use of the terminology also in the expression of expectations here. Hence, we subsumed the mentions of reproducibility with the mentions of replicability. Some respondents clarified the greater importance they attribute to replicability compared to criteria related to (re-)usability.

> Experiments should be reproducible. Good documentation and easy setup are a plus but we should keep in mind that an artifact should not be seen as commercial software. (id 41, 4 PL AECs)

Transcending the diversity found in submitted artifacts, replicability is the central criterion given for the acceptance or rejection of an artifact.

Good Documentation. Besides replicability, a well-prepared documentation of artifacts is also important for reviewers. In the context of the received responses, documentation can mean the description of the artifact and its parts, a description of setup procedures, or detailed documentation of individual parts (e.g., code comments). However, it does not seem to be a major reason for acceptance or rejection, as only five participants mentioned their decision to be influenced by documentation. It is also the first criterion listed for the *Artifact Evaluated – Functional* badge suggested by the ACM guidelines and the most prevalent criterion named in CfAs. However, it is not clearly specified what makes a *good* documentation, which is also mentioned as problematic by some participants.

> [...] Documentation is obviously fuzzier, but there at the bare minimum should be instructions that tell a reviewer how to run the artifact and reproduce said results. [...] (id 237, 2 PL AECs)

Easy Setup. The ease of the setup process for an artifact is also often mentioned as a minimum requirement. One participant explains: "It should not require more than 2 hours of effort on the part of the evaluator to kick off the results evaluation process." This finding is in line with our analysis of the ACM guidelines and CfAs in Section 3, where the actual time limit set by the calls is significantly shorter.

Further Insights. We found that some reviewers go beyond the replication of experiments from artifacts and also manipulate the experiments, which is encouraged by 24 of the analyzed 79 CfAs.

> I follow a process corresponding to badge criteria
> 1. I read the paper and check that mentioned artifacts exist.
> 2. I search for the Readme that describes the setup (or data). I evaluate based on clarity of the setup guide.
> 3. I search for provided demos and test cases or reproduction scripts.
> 4. I try to create a problem (a test case, such as a new language that is supposed to be implemented with provided tool) and solve it based on the artifacts provided. (id 163, 1 PL AEC)

Interestingly, we also found that some participants of the survey mentioned that they care more about artifact availability than for their quality. This was surprising to us because it would indicate that detailed quality criteria beyond the artifact supporting the claims made in the paper might be obsolete.

> [. . .] It's much more important that something is available than its quality. If the authors published a paper using this code/data/whatever, it would be good if the code/data/whatever was available – even if it's low quality. Enforcing quality criteria only means that some authors will not publish their code/data/whatever, but the paper is still published. (id 96, 1 SE AEC)

Expectations for Specific Artifact Types. While most CfAs name different types of research artifacts (i.e., code, proof, and data), they do not state different criteria for those. To assess if different criteria are used in practice, we asked our participants whether they have different expectations for different artifact types.

We received 123 answers regarding code artifact, 105 answers regarding proof artifacts, and 112 answers regarding data artifacts.

Code. Documentation in various forms was mentioned most often as quality criteria for code artifacts. Specifically, the participants mentioned *documentation in general* (30: 19/9), *setup documentation* (17: 16/1), *code documentation* (6: 5/–), *documentation only of relevant parts* (4: 3/2), *documentation of command-line options* (2: 1/1), and several specific single mentions (externally exposed features, file formats, usage) as important for code. In addition to documentation, *code should compile and run* when provided as an artifact, as 29 (23/5) participants stated.

Code quality seems to be a debated criterion, especially in the PL community: While 19 (16/–) participants explicitly mention code quality as a minimum expectation, 12 (11/1) participants see code quality as not important for acceptance.

> I generally have low expectations for code, since I think the community generally favors proof-of-concept code over production-quality code. [. . .] (id 193, 2 PL AECs)

Additionally, 3 (3/–) participants mentioned that during artifact evaluation there would be no time to inspect code quality and one participant mentions that authors would not have the time to document or improve quality.

Among other mentions are *packaging* (12: 12/2), *legible code* (10: 9/1), and *easy setup* (8: 5/3).

Thus, regarding code, we observe a general understanding that documentation in several forms is a minimum expectation for a code artifact. However, we see a moderation in the amount of documentation requested. While it is undebated that a code artifact should compile and run, we found that there are differing views on the importance of code quality.

Proofs. For proof artifacts, respondents named the following quality criteria: *understandability* (24: 16/6), *completeness* (23: 19/4), *proof checker ran without errors* (12: 11/2), and *correspondence between claims from the paper and the formalized lemma* (8: 6/1). Again, documentation in various forms is mentioned frequently: *documentation of the high-level flow* (9: 8/–), *documentation in general* (8: 6/1), *comments on definitions* (4: 4/–), *documentation on how to compare to paper results*, *documentation of any assumptions*, *documentation of usage beyond the paper.*

During artifact evaluation, proofs appear to be more rated for their internal properties, such as understandability or completeness, rather than on their ability to proof check without error, which was criticized by some respondents.

Data. For data artifacts, we found the following quality criteria: *format description* (33: 19/8), *raw data included* (16: 13/2), and *documentation in general* (13: 10/2). Further mentioned were *non-proprietary formats* (8: 6/4), *reproducibility* (8: 7/1), *completeness* (7: 4/2), and *script/program/library to manipulate data* (7: 7/–).

Thus, several participants expect that not only the data should be contained in the artifact, but also the scripts, programs, or libraries necessary to manipulate, analyze, or plot the data.

> The raw data of the original submission should be included + a script/tool to plot what is in the paper. Data might be correct or not but also the plotting can contain bugs disturbing the message.
> (id 156, 1 PL AEC)

Summary and Discussion. Considering the reviewer perspective, we found that replicability of results is the most important criterion for the acceptance or rejection of artifacts, which is in line with our analyses of criteria set forth by CfAs. This result is dominated by the larger group of responses we received from the PL community. Replicability is not mentioned in answers for specific artifact types, no matter from which community. Hence, we conclude that replicability is more a general property attributed to the whole artifact regardless of its type. If the results reported by the authors in their paper can be replicated, the artifact is generally considered of sufficient quality to be accepted for the artifact evaluation track.

As mentioned previously, there are two distinct views on the quality criteria for artifact evaluation. While the first perspective is that the availability of an artifact (cf. Section 4.2) is more important than its quality as long as it meets the expectations set by the paper, the other perspective is that the quality of accepted artifacts needs to improve beyond this. In the software engineering community, the creation of higher standards is visible for ICSE and FSE, as both conferences do not award the *Artifact Evaluated - Functional* badge anymore[9], but rather award either the *Reusable* or just the *Available* badge. In the respective CfAs, this is justified by the objective of the artifact evaluation track to foster reusable artifacts.

> We found that there is no consensus on the topic of a well-defined quality threshold. However, some conferences in the software engineering community established higher requirements for artifacts.

We found different expectations depending on the artifact type. Although documentation is mentioned for all three artifact types, especially for proof and code artifacts, there are different expectations, such that code requires documentation, and proofs require completeness and understandability. This is not surprising, because program code can be supplied in multiple forms and languages, whereas mechanized proofs are usually formulated using one of the major proof assistants (i.e., Coq, Isabelle, etc.). For proofs, the complexity here lies more in the formulation of the theory itself, which needs to be explained step by step, hence motivating the requirement of understandability.

> Reviewers expect different quality criteria for different artifacts types, but these are not communicated explicitly in CfAs.

4.3.2 *Perspective as User.*

To assess the expectations towards research artifacts from a user perspective, we asked the participants of our survey (1) how many artifacts they have used for other reasons than evaluating them, (2) whether their expectations toward any reused artifacts were met, and (3) to elaborate on their (un-)met expectations. A total of 128 participants completed this part of the survey. If the respondents replied how many artifacts they have used, we include this information along with the respondent ID in the quotations.

Quality Criteria/Expectations. Most positively, many participants were satisfied with the artifacts they (re-)used, irrespective of whether it was code (45 satisfied vs. 12 not satisfied), proofs (10/2) or data (25/7). This indicates that, whatever criteria are applied, the checks for reusability in artifact evaluation processes cover what is expected by (expert) users. With more than 20 % dissatisfaction, there is nonetheless clear room for improvement. Like for the reviewer perspective, the expectations differed for each artifact type.

Code. For code artifacts, the dominating quality criteria were *documentation* (14) and *runnability* (10). These were followed by *reusability* (7) and *result replicability* (4). Less frequently mentioned criteria were *usability* (2), *source code availability* (2), and *code quality* (2).

Regarding documentation, participants indicated different purposes: First, documentation should help to *explain how results can be replicated*: "Pure open source software repositories often lack the documentation, scripts and benchmark codes required to replicate a research paper. [. . .] we required the extensive help of the first author of a paper to be able to use it as comparison point in our own paper" (id 246). Second, documentation should *explain how code works*: "I was able to (1) see enough to get a sense of how to do it myself, and (2) easily determine that their implementation would not work for my purposes" (id 98, 1-5 code & 1-5 data artifacts). Third, documentation should *explain how it can be extended*: "I expected it to have enough documentation so that I understand where to put my extensions, and it did" (id 145, 1-5 code & 1-5 proof artifacts).

Runnability seems to be mostly perceived as a binary criterion, as participants reported that the code artifacts they used "ran" or "worked". Problems for code that does not run can be caused by lacking maintenance of both documentation or code after the initial submission and publication of the artifact.

> [. . .] Even when the code is useful and functional, the documentation is usually out of date. Most of the time, I spend a day or two trying to make it run, only to give up once I run into sufficiently hard problems. Other times the code is so outdated that there is no way to make it work without completely updating it. [. . .]
> (id 216, 10-20 code & 5-10 data artifacts)

Proofs. Only few participants indicated experience with using proof artifacts, and the few responses saw *understandability* (3) and *(re-)usability* (2) as important quality criteria. Understandability mainly covers aspects of how mechanized proofs correspond to claims in articles, whereas (re-)usability of proofs relates to artifact handling or reuse of parts from proof artifacts with other code.

Data. For data, *availability* was the most important quality criterion (5), followed by its relation to actual *raw data* (4). The availability of data in addition to result summaries commonly reported in space constrained research articles is perceived as valuable, but has to overcome limitations:

> [. . .] A couple of times papers have referred to publicly available datasets from other sources, that seem to have moved or disappeared since then. (id 216, 10-20 code & 5-10 data artifacts)

Another concern was how available data relates to raw data. One way to ensure traceability to raw data in data artifacts is to provide the raw data along with automated analyses, which have been explicitly mentioned as an important criterion for data artifact quality by some participants: "Sometimes data are aggregated and

[9] At FSE since 2018.

other cases it is not clear how to obtain the final results from the raw data."(id 39, 1-5 code & 5-10 data artifacts)

Summary and Discussion. Regarding artifact usage, the expectations vary for different artifact types. Documentation, as the most prominent concern for code artifacts, is rated even higher than the code's runnability, probably due to our expert respondents' confidence to get code to run if only the documentation is good enough. Similar to our results in Section 4.3.1, understandability is of high concern for proofs. For data artifacts, availability and raw data are of higher concern than documentation.

The expectations on code artifacts show a higher number of replies related to reusability (7) than to replicability (4). This is corroborated by open comments on artifact usage, in which 14 respondents indicate reusability as artifact purpose, whereas only 6 indicate replicability. We observe this prevalence of reusability over replicability despite a majority of 55 PL AEC members, for whom replicability dominates as AE purpose in CfAs, over 18 SE AEC members, for our free text questions on artifact usage.

While the specific quality criteria differ by artifact types, artifact users generally find reusability more often an important purpose for providing artifacts than replicability. Although this observation may seem unsurprising at first, it indicates that artifact users do not perceive replicability as a positive effect on reusability, even though the preparation of a replicable artifact does require a similar set of criteria (e.g. documentation).

Respondents did not perceive replicability as a beneficial factor to reusability.

The artifact users in our survey were generally satisfied with the quality of the artifacts they used. However, this satisfaction is not clearly attributed to the quality assurance that artifact evaluations provide. While 20 of the respondents indicated a notable difference between successfully evaluated artifacts and not evaluated artifacts, 34 indicated to not have observed such difference. The most frequently reported differences between successfully evaluated and other artifacts were *understandability*, *usability*, *consistency with paper results*, *availability*, and other less specific *quality* aspects.

To put our results on artifact usage in perspective, we need to point out that only 76 respondents have indicated to have any experience with artifact (re-)use beyond artifact evaluation. This needs consideration when interpreting our results, but also raises the question if artifact reuse is an uncommon scenario and, if so, why. While answering this question is beyond the scope of our study, we deem it important to report the observation as a result to be addressed by future research.

Despite the promotion of artifact reusability as a central goal of artifact evaluation in many CfAs, less than half of the respondents in our study reported to have experience with artifact (re-)use.

4.3.3 Discussion/Comparison Between Perspective of Reviewer and User. From the answers we collected we could see that there are very diverse expectations toward artifact quality among respondents. For the largest part, the expectations mentioned by respondents fall into larger categories that match the evaluation criteria stated in CfAs and in the ACM guidelines: Documentation, consistency, completeness, exercisability, and reusability. However, we find the *importance* of these criteria to differ for different types of artifacts and depending on whether the perspective of a reviewer or an artifact user is assumed.

While *replicability* is a criterion frequently mentioned from the reviewer perspective, from which it is a central criterion for artifact acceptance or rejection, it plays a much lesser role from a user perspective, which (unsurprisingly) favors reusability over replicability. Besides this difference, the reported quality criteria do not significantly differ, but the set of criteria stated for artifact review is more diverse. On the one hand, this observation gives confidence that criteria that are important for reusability are already adequately covered by existing artifact evaluation processes. On the other hand, our observation is based on few responses on artifact reuse, which mandates further investigation.

If regarded separately by artifact type, we find the various criteria to be of different importance. For code artifacts, documentation is the most important criterion and we received very detailed views on what is expected to be covered by documentation and to which degree of precision. While exercisability seems to be an obvious criterion that does not require further elaboration, code quality is less specific and discussed differently by respondents. This may lead to ambiguities in the review process and we would recommend AEC chairs to address this accordingly in future CfAs or review guidelines. For proofs, documentation is also considered very important, but less than understandability, which appears to be an equally unspecific term as code quality for code artifacts and we recommend clarification in the CfAs. For data artifacts, it is important to respondents that raw data and manipulation scripts are included in the artifact submission in addition to proper documentation, especially of formats used. The inclusion of raw data and scripts is a fairly unambiguous criterion that AEC chairs may want to consider to include in the submission criteria for data artifacts.

Reviewing vs. using artifacts elicit different expectations regarding quality. In both views, expectations toward artifacts vary for different artifact types, some of which lack clear definitions in the ACM guidelines and CfAs, which may lead to misunderstandings.

5 FURTHER INSIGHTS

Our study revealed insights beyond the expectations toward artifacts. In this section, we present and discuss these findings.

5.1 Satisfaction with the Evaluation Process

We were interested in the opinion of the reviewers toward the current practice of artifact evaluation and asked "Do you think that the effort of artifact evaluation is justified?". In general, we found that the effort for reviewers is perceived as justified. Specifically, our participants found that artifact evaluation guides authors toward good artifacts. We also found a few interesting cases, e.g., discovering fraudulent research:

> I once flagged a clearly fraudulent artifact. Its outputs were hard-coded into the source code. Not only was the artifact rejected, but the main PC was notified, the authors were contacted, and the paper withdrawn. This is a good outcome, having kept bad work out of a top conference. (id 193, 1 PL AEC)

However, when we asked for satisfaction with the current artifact evaluation process, the answer were mixed. Some respondents were satisfied, some saw potential to improve the process. For instance, one participant reported:

> I think there is room to improve but that the conferences are actively doing so. The process seems to be functioning.
> (id 184, 2 PL AECs)

Others were pointing toward various shortcomings such as reviewer attention or recognition, as the following participant reported:

> Some reviewers are thorough, some are not. [. . .] And AEC should be rewarded with better recognition.
> (id 61, 1 SE AEC, 2 PL AECs)

A frequent criticism (13 out of 66 answers) of the respondents was related to a missing quality standard for AE. Respondents stated they were missing clear criteria according to which artifacts are accepted/rejected or badges are awarded and that the ACM guidelines are too generic and open to interpretation to serve as a quality standard. As these interpretations can differ across conferences, the interpretation of what a badge really means can only be understood in the context of a given conference (and year).

Besides these difficulties regarding thresholds for the AE outcome, respondents criticized missing guidance *how* AE should be conducted, i.e., which steps should be taken or which criteria checked for, and suggested the development of checklists, "templates" and "benchmarks" for AE to provide guidance to AECs.

Although only half the reviewers are satisfied with the current artifact evaluation process, most of them still see that evaluation is worth it. Our findings indicate that with artifact evaluation, we are on the right track and we should continue. However, there is also room for improvement. In particular, a common quality standard for artifacts and common review guidelines need to be developed.

5.2 Reviewers' Experience with Artifacts

Artifact evaluation committees are usually recruited out of junior researchers. As the process has now been established for several years at major conferences, we were curious if reviewers themselves have experience in preparing and submitting artifacts as required for a true peer review. Out of 115 participants, 76 (66.1 %) indicated to have submitted a research artifact for evaluation before, i.e., a large group of reviewers has no experience *composing* a research artifact. Moreover, in 4.3.2 we reported that few reviewers have experience *(re-)using* a research artifact. As artifact evaluation is a comparatively new process, this was expected but leaves room for improvement. In particular, this means that CfAs, review instructions from AEC chairs, and opinions from fellow reviewers are the sole source of criteria according to which artifacts are evaluated for many AEC members. Given the lack of quality criteria and review guidance indicated by our respondents (cf. Section 5.1) and the high fluctuation of AEC members as indicated by the average number of committees served on (cf. Figure 2), the lack of reviewer experience currently puts an enormous impact (and responsibility) on how AEC chairs steer the review process.

5.3 Review as an Interactive Process

Several participants value close communication with the authors and would like to increase the interactivity in reviews beyond the currently established *kicking-the-tires phase* in many AE processes, where close communication is enabled for clarifying setup issues.

> [. . .] Ideally, I would like to see a two-step process of evaluating the artifact and submitting an improved. However, this increases the load on the artifact evaluation committee. (id 22)

While a multi-step process might be too difficult to realize for all artifacts, some artifacts may benefit from this way of *shepherding* much alike paper submissions. However, a *kicking-the-tires phase* or other interactive processes have only been implemented in 8 out of the 16 conferences in 2019 according to their CfAs.

5.4 Tighter Coupling to Paper Acceptance

13 participants indicated their support for a tighter coupling between artifact evaluation and paper acceptance although this was not part of any question. Suggestions range from shepherded acceptances, mandatory artifact submissions for specific tracks (e.g., as is the practice for the tool tracks for CAV and TACAS already), to having artifact submission mandatory for all paper submissions.

> We need to have a 'conditional acceptance' that depended on the result of the artefact evaluation: I'm not proposing this to be 'the norm', but a special case of acceptance like 'shepherding', where few papers are accepted only if the artefact withstand the statements of the paper. (id 214, 1 PL AEC)

If artifact evaluation does not have any influence on the acceptance of a paper, reviewers struggle with the incongruousness between artifact evaluation's mission statement of replicable research and the current practice. One participant reports:

> The evaluations do not determine if the paper is accepted or rejected, so in the bigger picture, I didn't much attention to these reviews.
> (id 79, 1 PL AEC)

In conclusion, the community suggests more rigor by integrating artifact evaluation much stronger into the paper review process.

6 THREATS TO VALIDITY

6.1 Internal Validity

The recollection of our participants' experiences with artifacts can be affected by the time span between their occurrence and the participation in our survey. Moreover, the perceived purpose of artifacts and the processes of their assessment have changed over time. Figure 1 shows that AEC members from every year since the initiation of artifact evaluations have participated and most participants have served in recent years. We, therefore, do not expect effects of time to significantly affect our results.

The second threat is related to our participant selection. To receive opinions from researchers familiar with artifact evaluation, we only invited AEC members to our survey. Their responses may be affected by concerns about the perceived value of their work, which can lead to overly positive reports regarding accepted artifacts or overly negative reports regarding rejected or not evaluated artifacts. We addressed this by openly communicating the anonymization policy of our study. Moreover, the main results we report are related to the expectations toward artifacts rather than the positive or negative experiences they have led to. Consequently, we expect the central conclusions to not be affected by this threat.

To assess the effectiveness of our instruments, we conducted pre-testing with 6 participants. If this sample is not representative of how the targeted audience perceives our questions, this can induce systematic effects in our results. Two participants in the pre-test had no experience with artifact evaluations and commented on the first draft version of our survey. The remaining four pre-tests have been conducted by experienced researchers. While we cannot rule out effects on our results in principle, most of the pre-testers' comments were in line and we have not seen symptoms of severe misunderstandings in the replies beyond what is caused by differing understandings of participants (which we intended to capture).

6.2 External Validity

A threat to the external validity of our study lies in the selection of participants. Our goal is to assess the community's expectation toward artifacts. With our focus on AEC members, we potentially create a bias toward specific expectations. Artifact evaluations are a relatively new scientific peer review process and, as such, a minority of researchers have working experience with artifacts or knowledge of the ACM guidelines. We, therefore, gave preference to the risk of selection bias over the risk of our questions being misunderstood.

Three of our pre-testers received invitations to participate in the survey and their replies may have been affected by the pre-test. We deem this risk tolerable, as the influence of three responses is marginal to the presented results from 257 replies.

7 IMPLICATIONS

In our view, artifact evaluation chairs, steering committees, and community leaders can take several actions to address the issues highlighted in this paper.

First, as a community, we have to define the purpose of artifact evaluation clearer than we do now. As our discussion showed, a twofold purpose of replicability and reusability has several pitfalls even though the two goals share certain characteristics. A committee might be instated for each community that defines a clear goal specific to the respective community. This committee should evaluate changes over the course of time and observe the improvements made (e.g., in terms of more submissions to artifact tracks).

Second, we propose to also work on agreed quality standards in each community in a similar manner. As we have shown, expected artifact quality criteria vary widely between communities, perspectives, and artifact types. Moreover, there is a need to communicate quality criteria very clearly. Misconceptions about appropriate acceptance levels seem to be common during artifact evaluation and can lead to serious conflicts as one respondent pointed out.

> Two students argued for acceptance of the artifact because it was capable of generating output without crashing in some scenarios. I argued strongly against them and the artifact was eventually rejected. [...] (id 265, 1 SE AEC)

As most reviewers only serve once on an AEC (cf. Figure 2) chairs should explicitly brief AEC members on appropriate acceptance levels and quality criteria. It would not be beneficial to push these specific criteria into a discipline-wide document such as the ACM guidelines. Rather, CfAs should be extended to incorporate those criteria. As it seems to be common practice that chairs "inherit" those CfAs from their predecessors, the quality criteria can evolve over time and reflect community transitions in a fine-grained way that is also trackable for evaluations such as the one presented here. However, community representatives should monitor if a common core can be established within a community, which we consider most likely from our results.

Third, recruiting reviewers also based on their experience as artifact creators may not only benefit the quality and efficiency of artifact reviews, but also improve peer consultation for new and less experienced reviewers.

Furthermore, artifact reviewers clearly felt that the time has come for a tighter coupling between artifact evaluation and paper acceptance. Conference steering committees and track chairs should take the opportunity to incorporate artifact evaluation into acceptance processes. Tool-oriented tracks could take the lead (as TACAS and CAV show) and other tracks could follow closely learning from the experience gained in the past decade.

8 CONCLUSION

The replicability crisis shook the research community, and also reached the software engineering and programming language community. A recent attempt to mitigate its rippling through the communities has manifested in the creation of artifact tracks at conferences. However, at this point it is still unclear if or to which degree artifact evaluations are a suitable measure to foster replicable, let alone reproducible, research. Taking a first step toward an assessment of artifact evaluations, we inspected how the process is currently seen by the community, specifically by the people who perform these evaluations. We found that the initial mission of artifact evaluation of assessing replicability has now grown to also cover reusability. However, reviewers and users of artifacts see a different purpose of artifacts: Reviewers want replicability, users want reusability. Additionally, different artifact types elicit different expectations, but they all more or less focus on making artifacts understandable and usable. Now, we as the research community need to clearly communicate these expectations in calls and guidelines while carefully defining and using terminology to avoid misunderstandings or false expectations. This is one approach to avoid a replicability crisis of the same large extent as it washed over the psychology community.

ACKNOWLEDGMENTS

We are very grateful to all members of the community that took time out of their busy days and replied to our survey. They provided us with detailed information and diverse views on the current state of artifact evaluation processes. Oliver Schwahn, Habib Saissi, Lisa Nguyen Quang Do, Norbert Siegmund, and two other colleagues provided us with valuable feedback during the pre-test of our survey. We thank the anonymous reviewers for their helpful comments which improved this paper We greatly appreciated inputs from Jon Bell and Tim Menzies that significantly improved the final version of this paper.

REFERENCES

[1] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

[2] Anna Balazs. 2008. International vocabulary of metrology-basic and general concepts and associated terms. *Chemistry International* (2008), 20–1. https://doi.org/10.1515/ci.2008.30.6.21

[3] Victor Basili, Forrest Shull, and Filippo Lanubile. 1999. Building Knowledge through Families of Experiments. *IEEE Trans. Softw. Eng.* 25, 4 (1999), 456–473. https://doi.org/10.1109/32.799939

[4] Emery D. Berger, Celeste Hollenbeck, Petr Maj, Olga Vitek, and Jan Vitek. 2019. On the Impact of Programming Languages on Code Quality: A Reproduction Study. *ACM Trans. Program. Lang. Syst.* 41, 4, Article 21 (Oct. 2019), 24 pages. https://doi.org/10.1145/3340571

[5] Karl Broman, Mine Cetinkaya-Rundel, Amy Nussbaum, Christopher Paciorek, Roger Peng, Daniel Turek, and Hadley Wickham. 2017. Recommendations to funding agencies for supporting reproducible research. https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf. Accessed: 2020-09-03.

[6] B. R. Childers and P. K. Chrysanthis. 2017. Artifact Evaluation: Is It a Real Incentive?. In *2017 IEEE 13th International Conference on e-Science (e-Science)*. 488–489. https://doi.org/10.1109/eScience.2017.79

[7] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in Computer Systems Research. *Commun. ACM* 59, 3 (Feb. 2016), 62–69. https://doi.org/10.1145/2812803

[8] Erin Dahlgren. 2019. Getting Research Software to Work: A Case Study on Artifact Evaluation for OOPSLA 2019. https://doi.org/10.5281/zenodo.4016657

[9] Association for Computing Machinery. 2018. Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging. Accessed: 2020-09-03.

[10] Leonid Glanz, Sven Amann, Michael Eichberg, Michael Reif, Ben Hermann, Johannes Lerch, and Mira Mezini. 2017. CodeMatch: Obfuscation Won't Conceal Your Repackaged App. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017)*. Association for Computing Machinery, New York, NY, USA, 638–648. https://doi.org/10.1145/3106237.3106305

[11] Matthias Hauswirth. [n.d.]. Artifact Evaluation. http://evaluate.inf.usi.ch/artifacts. Accessed 2020-09-03.

[12] Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. *Community Expectations for Research Artifacts and Evaluation Processes - Data & Scripts*. https://doi.org/10.5281/zenodo.3951724

[13] Robert Heumüller, Sebastian Nielebock, Jacob Krüger, and Frank Ortmeier. 2020. Publish or Perish, but do not Forget your Software Artifacts. *Empirical Software Engineering* (2020). https://doi.org/10.1007/s10664-020-09851-6 Preprint.

[14] William Hudson. 2013. Card Sorting. In *The Encyclopedia of Human-Computer Interaction*. The Interaction Design Foundation, Chapter 22.

[15] Natalia Juristo and Sira Vegas. 2011. The Role of Non-exact Replications in Software Engineering Experiments. *Empirical Software Engineering* 16, 3 (2011), 295–324. https://doi.org/10.1007/s10664-010-9141-9

[16] Shriram Krishnamurthi and Jan Vitek. 2015. The Real Software Crisis: Repeatability As a Core Value. *Commun. ACM* 58, 3 (Feb. 2015), 34–36. https://doi.org/10.1145/2658987

[17] J. Lung, J. Aranda, S. Easterbrook, and G. Wilson. 2008. On the difficulty of replicating human subjects studies in software engineering. In *2008 ACM/IEEE 30th International Conference on Software Engineering*. 191–200. https://doi.org/10.1145/1368088.1368115

[18] Daniel Méndez Fernández, Wolfgang Böhm, Andreas Vogelsang, Jakob Mund, Manfred Broy, Marco Kuhrmann, and Thorsten Weyer. 2019. Artefacts in software engineering: a fundamental positioning. *Software & Systems Modeling* 18, 5 (2019), 2777–2786.

[19] Daniel Méndez Fernández, Martin Monperrus, Robert Feldt, and Thomas Zimmermann. 2019. The open science initiative of the Empirical Software Engineering journal. *Empirical Software Engineering* 24, 3 (2019), 1057–1060. https://doi.org/10.1007/s10664-019-09712-x

[20] Gregorio Robles. 2010. Replicating MSR: A study of the potential replicability of papers published in the Mining Software Repositories proceedings. In *2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010)*. 171–180. https://doi.org/10.1109/MSR.2010.5463348

[21] Forrest J Shull, Jeffrey C Carver, Sira Vegas, and Natalia Juristo. 2008. The role of replications in Empirical Software Engineering. *Empirical Software Engineering* 13, 2 (2008), 211–218. https://doi.org/10.1007/s10664-008-9060-1

[22] Janet Siegmund, Norbert Siegmund, and Sven Apel. 2015. Views on Internal and External Validity in Empirical Software Engineering. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. 9–19. https://doi.org/10.1109/ICSE.2015.24

[23] Christopher S. Timperley, Lauren Herckis, Claire Le Goues, and Michael Hilton. 2020. Understanding and Improving Artifact Sharing in Software Engineering Research. arXiv:cs.SE/2008.01046

[24] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376448