Survey-Based Insights into the Replication Crisis and the 3R in Software Engineering

Ivanildo Azevedo
vando.aze@gmail.com
Universidade Federal de Pernambuco
Brazil

Ana Paula Vasconcelos
aninha.lfv@gmail.com
Universidade Federal de Pernambuco
Universidade do Estado de Mato Grosso
Brazil

Eudis Teixeira
eudis.oliveira@ifsertao-pe.edu.br
Instituto Federal do Sertão Pernambucano
Brazil

Sergio Soares
scbs@cin.ufpe.br
Universidade Federal de Pernambuco
Brazil

Abstract

Context: Efforts to improve reproducibility and research credibility have gained relevance in multiple fields, including Software Engineering (SE), where the 3R practices (Repeatability, Reproducibility, and Replicability) are essential to ensure the reliability of empirical studies. Despite growing interest in Open Science, concerns about a Replication Crisis persist.

Objectives: To assess the perceptions of SE researchers of the Replication Crisis and 3R practices, identify good practices, barriers, and facilitators to reproducible research, and evaluate the community's acceptance of the ACM's standardized definitions of Repeatability, Reproducibility, and Replicability.

Method: We conducted a survey adapted from Baker [5], targeting authors of SE studies related to replication. From a list of 1,061 researchers, we received 101 responses. The questionnaire combined Likert-scale and open-ended questions. Responses were analyzed using descriptive statistics and Reflexive Thematic Analysis.

Results: Most respondents (94.1%) acknowledged the importance of 3R practices. Although on average 84.5% agreed with the ACM definitions, participants raised concerns about their clarity and applicability, especially to qualitative research. The majority (74.3%) recognized the existence of a Replication Crisis in SE. The key challenges reported include a lack of protocols, selective reporting, data unavailability, and pressure to publish. Positive actions included using containers (e.g., Docker), version control, artifact sharing, and Open Science practices. However, participants noted that cultural and institutional incentives for reproducibility remain limited.

Conclusion: Although SE researchers support the principles of 3R practices and recognize ongoing challenges, uncertainty persists about the scope and solutions of the crisis. This study highlights the need for more precise terminology, better reporting standards, and greater institutional support to promote reproducibility, transparency, and research integrity in SE.

Keywords

Replication Crisis, Repeatability, Reproducibility, Replicability, 3R, Software Engineering, Open Science

1 Introduction

Several studies from the most diverse disciplines indicate high failure rates in attempts to replicate/reproduce research [5, 9, 15, 22, 26]. This problem is known as a Replication Crisis "in which experimental results cannot be reproduced and published findings are mistrusted" [15]. This crisis was initially observed in behavioral, cognitive, and social disciplines [34]. However, it also occurs in Computer Science fields, such as Artificial Intelligence [21], Information Systems, and Software Engineering [15], raising serious concerns about the reliability and validity of conclusions drawn in several studies [34]. Reliance on undocumented or unavailable tools in SE impedes replication [29]. Clear guidelines, transparency, and open data sharing are critical to improving replicability and reliability [20]. As a result, addressing reproducibility issues has become a critical concern, demanding greater transparency, artifact availability, and methodological rigor.

3R practices are essential for reliable research findings, but have not been sufficiently explored in SE [30]. Insufficient data disclosure by original authors often hampers replication efforts, casting doubt on the reliability of published results [15]. Additionally, the absence of widely adopted replication guidelines exacerbates the issue [3, 30]. The confusion between the 3R terms, such as replicability and reproducibility, has been reported in several fields [3, 16, 19, 20, 25, 28], and further complicates the situation, leading to inconsistencies in how research results are communicated and interpreted. These concepts are fundamental to validating and ensuring the reliability of scientific results. Although initiatives aim to align terminology [1] with broader scientific communities, variations in definitions persist, often leading to divergent interpretations, where researchers employ different definitions that can have significantly different meanings.

The terms **Repeat**, **Reproduce**, and **Replicate** refer to the specific acts of duplicating a study. **Repetition**, **Reproduction**, and **Replication** refer to the broader processes of conducting such duplications, while **Repeatability**, **Reproducibility**, and **Replicability** describe the extent to which the results of a study can be consistently duplicated. Furthermore, **Repeatable** indicates that consistent results can be obtained when the study is repeated under identical conditions by the same research team, **Reproducible** means that the results can be achieved using the same data and analysis methods in the same or different context by a different research team, and **Replicable** suggests that the results of the study can be duplicated by a different research team redoing the original study with different conditions [1].

Previous studies have documented the Replication Crisis in several fields (such as Medicine [9, 22], Psychology [26], and Computer Science [15]), highlighting difficulties in reproducing or replicating results and questioning the reliability of published findings. In Software Engineering, existing research has investigated replication challenges and proposed strategies to address them [17], but often with a different focus, such as methodological guidance or insights from adjacent domains such as High-Performance Computing [4] or Synthetic Aperture Radar [7]. Other works have emphasized the confusion caused by inconsistent definitions of Replication, Reproduction, and Repetition (3R) [8, 20, 27, 28], which has hindered standardization efforts across disciplines [23]. However, to our knowledge, no previous study has investigated the Replication Crisis in SE while also examining the community's understanding and adoption of the Association for Computing Machinery (ACM) standardized 3R definitions.

We provide results for the following Research Questions (RQs):

• **RQ1 [3R]** *Do Software Engineering researchers agree with the terms Repeatability, Reproducibility, and Replicability?*

• **RQ2 [Replication Crisis]** *In the researchers' opinion, is there a Replication Crisis in Software Engineering?*

  – **RQ2.1 [Replication Crisis factors]** *What are the contributing and mitigating factors of the Replication Crisis in Software Engineering?*

To answer these RQs, we surveyed SE researchers to understand their perceptions of the Replication Crisis, challenges, and good practices. Our work offers insights into how experienced SE researchers perceive the Replication Crisis and the degree to which they align with the current concepts of Repeatability, Reproducibility, and Replicability (3R). We analyzed their perceptions on the Replication Crisis and the 3R terminology to describe different dimensions of research replication in Software Engineering, as defined by the ACM [1]. Our contributions are summarized as follows:

- **First in-depth investigation of the Replication Crisis in Software Engineering.** To the best of our knowledge, this is the first study to explore the topic with this level of depth, specifically in SE.
- **Evaluation of the Software Engineering community's alignment with ACM's 3R definitions.** While broadly accepted, respondents raised important concerns about clarity, scope, and applicability, especially for qualitative research.
- **Mapping of actual practices adopted by researchers to support 3R.** Highlights the recent growth in the adoption of practices, especially documentation and artifact sharing.
- **Identification of key contributors to 3R failures.** Missing data/code, academic pressure, and insufficient documentation were the most frequently cited factors.
- **Analysis of barriers when trying to reproduce or replicate studies.** Includes technical limitations, lack of transparency, and lack of support from original authors.
- **Insights into the cultural and institutional shifts needed to mitigate the Replication Crisis.** Suggests revisiting productivity metrics, prioritizing quality, and fostering supportive replication communities.
- **Categorization of community-driven suggestions to improve 3R.** Recommendations encompass cultural change (e.g., greater acceptance of replication studies and negative results), institutional reform (e.g., funding and recognition mechanisms), and technical support (e.g., tools, documentation standards, and infrastructure).
- **Commitment to Open Science through full artifact availability.** All research artifacts used in this study are publicly available (Section Artifacts Availability), promoting transparency, allowing verification, and supporting future research efforts.

The remainder of the paper is structured as follows. **Section 2** discusses related work, **Section 3** presents the methodology adopted in this study, **Section 4** presents the results and provides answers to our research questions, **Section 5** outlines our conclusions and addresses threats to validity, and the **Artifacts Availability** section details the disclosed data and provides a link to our repository.

2 Related work

Replication Crisis. Initially identified in health-related fields such as medicine [22], psychology [26], and cancer biology [9], it has also drawn attention in Computer Science [15]. Baker [5] found that more than 70% of the scientists struggled to reproduce others' experiments, and over 50% also faced challenges replicating their own work, prompting investigations in all disciplines, including SE. Our research adapts Baker's survey to focus on SE, incorporating questions about the 3R (Replication, Reproducibility, and Repetition) as contributing factors to the crisis. Balz and Rocca [7] studied reproducibility of Synthetic Aperture Radar (SAR) through a survey modeled after Baker's, finding that 75% of respondents faced replication issues and proposed solutions such as automated record keeping, version control, and containerization. While their focus was on SAR, our study examines SE and the factors that affect the 3R. Dos Santos et al. [17] explored replication challenges in SE, offering recommendations to researchers and practitioners, while our work emphasizes the Replication Crisis and guidance for conducting 3R studies. Finally, very recent replications of Baker's survey in Biomedicine [14] and in Sport and Exercise Science [24] report patterns that closely mirror our own results. These parallels are discussed in the Results section.

3R. Magalhães and Silva [23] interviewed SE researchers to develop a taxonomy for empirical study replication to promote discussion of standardized definitions. Our study diverges by assessing SE researchers' acceptance of ACM's 3R definitions. Antunes and Hill [4] examined the reproducibility crisis in High-Performance Computing (HPC), identifying how evolving processes obscure replication. Our work parallels theirs, but it focuses on SE through survey data. Studies such as Peng [27] and Barba [8] found that inconsistent 3R definitions cause confusion. Before 2020, ACM's definitions[2] differed from current ones [1], leading to inconsistencies. Barba [8] emphasized that many articles still use outdated ACM definitions, complicating standardization.

Novelty. To advance the discourse on research rigor within Software Engineering, this study investigates perceptions surrounding the Replication Crisis and the 3R, specifically how researchers conceptualize, engage with, and are affected by related challenges and

practices. Despite the growing body of literature on the 3R in various scientific domains, to the best of our knowledge, this is the first domain-specific investigation of the Replication Crisis and the 3R in Software Engineering that explicitly employs the ACM's standardized 3R definitions as a framework for analysis.

3 Study Design

Below, we present how we answered our RQs (presented in Introduction) and achieve our goal.

3.1 Data collection

We conducted a survey adapted for the Software Engineering context based on the original study by Baker [5]. Several questions from the original survey were retained. To address our research questions, we slightly modified some questions and added new ones. The final questionnaire, consisting of 31 items, was hosted on Google Forms, and collected anonymous responses between October 2023 and February 2024 to assess the participants' agreement with statements regarding the Replication Crisis and the 3R concepts. The complete questionnaire is publicly available in our research repository (see Artifacts Availability). Following the criteria established by Baldassarre et al. [6], this study qualifies as a conceptual replication of the original work.

Survey Pilot. To assess the clarity and comprehensibility of our questionnaire, we conducted two phases of tests involving six researchers from the Software Engineering research community. The questionnaire administered to the participants was the result of iterative refinements developed during both pilot phases.

Participants. We looked for experienced SE researchers with previous work on replication. Based on Systematic Mapping Studies of replication in Software Engineering, we identified papers that address replication[1] and extracted their authors to invite them to participate in our survey.

The Systematic Mapping Studies (SMS). We based our selection on the results of the SMS conducted by Silva et al. [31], which covered studies from 1994 to 2010, and by Bezerra et al. [10], who updated and extended the previous study [31] to include studies from 2011 to 2012. To ensure coverage of more recent publications, we further extended this mapping, adopting the research protocol (with some adjustments) and the refined search string[2] used by Bezerra et al. [10].

Two authors performed automated searches in the ACM Digital Library, IEEE Xplore, ScienceDirect, SCOPUS, and Springer. Additionally, we manually searched the proceedings of leading[3] journals and conferences in Software Engineering: Empirical Software Engineering Journal (ESEJ), Symposium on Empirical Software Engineering and Measurement (ESEM), Evaluation and Assessment in Software Engineering (EASE), International Conference on Software Engineering (ICSE), Journal of Systems and Software (JSS) and Transactions on Software Engineering (TSE). We also manually reviewed all available editions of two venues specifically focused on replication: the ROSE Festival (Rewarding Open Science Replication and Reproduction in SE) and the Replications Track of the International Conference on Program Comprehension (ICPC).

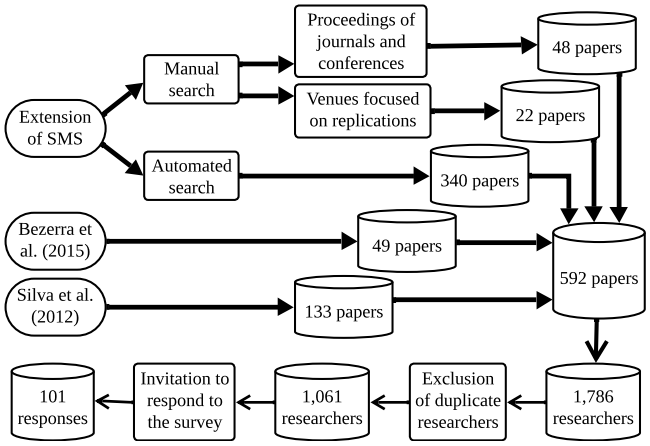

Figure 1: Process of identifying survey participants.

As presented in Figure 1, we retrieved a total of 592 papers: 133 papers from Silva et al. [31], 49 papers from Bezerra et al. [10], and 410 from our SMS extension. From these papers, we extracted the names and email addresses of 1,786 researchers. After removing duplicate researchers, we obtained a final list of 1,061 unique researchers. We individually invited the authors, via email, to participate in our survey. To encourage participation, a gentle reminder was sent two weeks after the initial invitation. We received 101 responses, which resulted in a response rate of approximately 9.5%. While typical in academic surveys, this rate may limit the generalizability of our findings due to potential non-response bias.

3.2 Data Analysis

The closed-ended questions, which assessed participants' level of agreement with specific statements, were answered using a five-point Likert scale. The analysis of these questions involved calculating the frequency of each response option to identify trends and general perceptions among participants.

For open-ended questions, we conducted Reflexive Thematic Analysis [12, 13]. The labeling process followed a six-phase process: (1) each researcher read all responses to familiarize with the data; (2) each researcher generated their initial labels for each response, after finishing, they met to analyze individual labels and reach a consensus to merge them; (3) labels were grouped into broader categories based on their similarities; (4) labels and categories were checked to make sure they agree with responses; (5) labels and categories were refined; and (6) the results were reported. As a result, a single response could receive multiple labels. This categorization process enabled us to identify patterns and recurring themes within the qualitative responses.

4 Results and discussion

Participants from several continents completed the survey. The majority are Ph.D. (89.1%), Professor Researchers (70.3%), have more than six years of experience performing empirical research

[1]Papers that study replication or which report a replication of a previous study.
[2]Bezerra et al. [10] refined the original search string "*to increase precision of the search process*", but confirmed that the revised version maintained the same sensitivity.
[3]According to [11, 18, 32, 33].

(73.3%), and are currently engaged in primary research (95.0%). The profile of the respondents demonstrates that we have mature and experienced researchers, providing us with more reliable results regarding the focus of our research.

4.1 3R (RQ1)

The goal of this research question is to examine the extent to which researchers accept and agree with the definitions of the 3R terms, assessing their clarity, applicability, and relevance within the Software Engineering research community.

Overall Agreement with ACM 3R Definitions

Among the ACM 3R definitions (Figure 2) Repeatability received the highest agreement[4] (89.1%), while Replicability had the lowest (78.3%), yielding an average consensus of 84.5%. However, 23 participants disagree[5] about the definitions, citing insufficient comprehensiveness for qualitative research, lack of clarity, and inadequate detail. Some criticized irrelevant aspects such as location or team, suggesting instead a focus on metrics such as the number of attempts required or the availability of links to the original study. Three participants proposed swapping definitions, arguing that Repeatability better describes Replicability, while others noted overlaps among the terms. Even those who agreed with the definitions highlighted the challenges in maintaining identical conditions, disagreements about defining aspects, and mismatches between definitions and real-world practices, reflecting the difficulty of applying 3R concepts effectively.

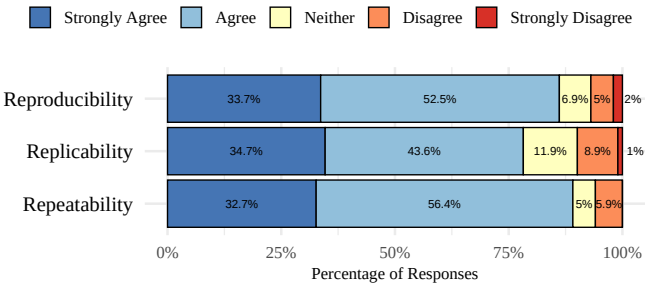

Figure 2: Agreement with the 3R definitions.

> **Takeaway 1:** Most of the participants agreed with the ACM definitions of Repeatability, Reproducibility, and Replicability, especially Repeatability. However, concerns were raised about clarity, applicability to qualitative research, and conceptual overlaps, revealing that the definitions may not fully reflect real-world research practices.

Concerns and Disagreements with 3R Definitions

Participants who expressed neutrality or disagreement[6] with the 3R definitions raised several concerns, many of which were common across the three concepts. The most frequent criticisms referred to a perceived lack of clarity and precision in the definitions, particularly regarding their applicability to qualitative research and studies involving human subjects. Respondents also questioned the relevance of certain elements (such as team composition or location) as defining factors, suggesting that these aspects may not always be meaningful in the context of Software Engineering.

Additional concerns included the omission of context-sensitive elements (e.g. how many repetitions are necessary for an experiment to be considered repeatable) and the narrow scope of the definitions, which seemed more suited to quantitative approaches. Two participants rejected all three definitions, arguing that they are overly aligned with quantitative research and overlook important nuances relevant to qualitative approaches.

> **Takeaway 2:** Respondents who disagreed or remained neutral cited lack of clarity, narrow applicability (especially to qualitative research), and questionable relevance of some defining aspects (e.g. location, team). Some felt that the definitions were overly focused on quantitative research, neglecting key contextual and methodological nuances.

Terminological Ambiguities and the ACM Revision

Some participants highlighted conceptual overlap between the definitions, particularly between reproducibility and replicability, others explicitly suggested revising or even swapping the two terms to better reflect their understanding. This issue is not unique to our participants and has been widely debated across various disciplines. Goodman et al. [19] pointed out that foundational terms, such as reproducibility and replicability, lack standardization, leading to conceptual and operational confusion on how scientific findings should be confirmed. Plesser [28] emphasized that common uses of reproducing and replicating often contradict long-established terminology in the experimental sciences. Cohen et al. [16] traced the lack of consensus to the fact that Reproducibility and Repeatability are nearly synonymous in general English and are frequently substituted for each other in formal definitions. In the context of Software Engineering, Hermann et al. [20] observed inconsistent, though often subtle, uses of replicability and reproducibility, attributing this partly to the evolving terminology and partly to discrepancies in the literature. Nosek and Errington [25] argued that the commonly accepted notion of replication/repetition as a procedure to observe if the same finding recurs is intuitive and widespread, yet inaccurate. They noted that vague theories and poorly understood methodologies often lead to misinterpretations of replication. Anchundia and Fonseca [3] also found that replication and reproducibility are frequently conflated in the literature, with varying definitions, some referring to replication as the extent to which a study matches the original, while others treat it as a synonym for reproduction or reanalysis.

A similar concern emerged within the Software Engineering research community itself when ACM introduced its artifact review and badge initiative and defined the 3R [2]. After discussions with the National Information Standards Organization (NISO), ACM received recommendations to align its terminology with the broader scientific community. Consequently, ACM decided to swap the definitions of Reproducibility and Replicability and retroactively updated all previously assigned badges to ensure consistency [1].

To mitigate these issues, it is essential that the 3R definitions be reviewed, refined, and clearly communicated, particularly with

[4]Participants who selected the options: "Agree" or "Strongly Agree".
[5]Participants who selected the options: "Disagree" or "Strongly Disagree".
[6]Participants who selected the neutral option ("Neither Agree nor Disagree") or expressed disagreement by selecting "'Disagree" or "Strongly Disagree".

regard to their applicability across different research paradigms. Greater clarity and precision would help reduce conceptual overlap and support greater understanding and consistent usage. This is especially important given that the terms Replication and Reproduction are still frequently used interchangeably in the literature, even within the same paper, as though they refer to the same concept. The feedback gathered in our study highlights an ongoing need for clearer guidance and shared understanding, particularly in the context of Software Engineering, where methodological diversity demands more nuanced and inclusive definitions.

Takeaway 3: Participants echoed long-standing debates across disciplines about the confusion between Reproducibility and Replicability. Several suggested swapping or revising the definitions. ACM itself revised its terminology in 2020 to align with broader scientific standards, highlighting the importance of refining and clearly communicating these definitions for consistency in the SE research community.

RQ1: Do Software Engineering researchers agree with the terms Repeatability, Reproducibility, and Replicability?
Our findings indicate that most Software Engineering researchers agree with the ACM definitions of the 3R, especially for Repeatability, which received the highest agreement level (89.1%). However, the results also reveal concerns and disagreements. Some participants felt that the definitions lack clarity, are too narrowly focused on quantitative research, or are not easily applicable to qualitative studies and human-centric research. Others pointed to confusion caused by overlapping terms, and a few even suggested that the definitions of Reproducibility and Replicability should be swapped. These perceptions align with ongoing debates across scientific fields, reflecting inconsistencies and a lack of consensus on how 3R concepts are understood and applied. Therefore, while general agreement exists, it remains a clear need to revisit and refine the definitions to ensure that they are inclusive, precise, and relevant across different research paradigms within Software Engineering.

4.2 Replication Crisis (RQ2)

Just under a decade ago, Baker [5] observed that 90% of researchers in various fields[7] believe that there is a Replication Crisis. In our survey, 74.3% of the respondents acknowledged the existence of such a crisis in Software Engineering, while 21.8% were unsure and 4.0% disagreed. Although our rate is slightly lower than that reported by Baker, it closely aligns with more recent findings in other research domains, such as Biomedicine (published in 2024) [14] and Sports and Exercise Science (published in 2023) [24]. A Murphy et al. [24] reported that 78.1% of Sports and Exercise Science researchers believed there was a Replication Crisis in their field, while Cobey et al. [14] found that 72% of biomedical researchers agreed that such a crisis existed in Biomedicine. These findings indicate that the perception of a Replication Crisis in Software Engineering is in line with other disciplines and sets the stage for a deeper examination of the specific challenges researchers face when attempting to Repeat, Reproduce, or Replicate prior studies.

This similarity in recent figures prompts a reflection: Why are the numbers today consistently lower than the original findings of Baker [5]? It is possible that strong evidence of the Replication Crisis over the years has become less prominent over the past decade, perhaps due to improvements in scientific practices. Or perhaps the scientific community has become desensitized to the issue through constant exposure, leading to a gradual normalization, and consequently broader acceptance of replication-related problems as an inherent part of research culture. The following section further explores the specific factors that contribute to this perceived crisis and those that may help mitigate it.

Takeaway 4: Most SE researchers perceive a Replication Crisis in their field, a figure consistent with recent findings in other scientific domains. Although slightly lower than earlier estimates in the literature, this trend suggests either gradual improvements in research practices or a growing normalization of replication challenges as part of the scientific process.

RQ2: In the researchers' opinion, is there a Replication Crisis in Software Engineering?
In the researchers' opinion, there is indeed a Replication Crisis in Software Engineering. A significant proportion of participants (74.3%) acknowledged its existence, a perception that is consistent with recent findings in other fields such as Biomedicine [14] and Sports and Exercise Science [24]. Although slightly lower than the original rate reported by Baker [5] in various disciplines, this alignment with more current studies suggests that the issue remains salient. These results support the notion that concerns about the reproducibility of scientific research are not unique to Software Engineering but reflect a broader, ongoing challenge across research domains.

4.3 Replication Crisis factors (RQ2.1)

To better understand the roots of the Replication Crisis in Software Engineering, we investigated the factors that researchers believe contribute to it. This analysis combines the perceptions of researchers with the practices they adopt (or neglect) to support them. The findings presented below provide a multifaceted view of both the obstacles and the efforts made to strengthen research Reproducibility, Repeatability, and Replicability.

3R as a significant challenge
Most of the participants (85.1%) agree[4] that Repeatability, Reproducibility, and Replicability (3R) in scientific research is a significant problem in the Software Engineering research community. When asked why (see Table 1) they pointed to several contributing factors: the unavailability or inadequacy of necessary resources (R2), issues inherent to the field of Software Engineering (R3), challenges related to the research culture and academic system that discourage replication efforts (R4), and the lack of clear, accepted, and practical guidelines for conducting and reporting reproducible studies (R5). Furthermore, several respondents reported experiences with unsuccessful attempts to reproduce or replicate studies (R6) or expressed personal reflections, either supporting or questioning the relevance

[7] Such as Chemistry, Physics and Engineering, Earth and Environment, Biology, and Medicine.

of the 3R in research (R1). These factors make replication and reproducibility more challenging to achieve consistently, hindering the feasibility of reproducing or replicating studies and thereby reinforcing the perception of a Replication Crisis.

Table 1: Reasons for considering the 3R a significant problem.

ID	Category	Examples	Total (52)
R1	Individual Perceptions and Value Judgments	e.g., positions of support, skepticism, or neutrality.	20 (38.5%)
R2	Lack of Technical or Material Support	e.g., unavailability or inadequacy of necessary resources, such as artifacts, data, statistical rigor, or detailed methodological descriptions.	14 (26.9%)
R3	Domain-Specific Constraints in SE	e.g., variability in study contexts, human factors, evolving technologies, and experimental complexity.	12 (23.1%)
R4	Structural and Process-Related Issues	e.g., lack of incentives, pressure to publish novel results, and the undervaluation or underuse of replication studies.	7 (13.5%)
R5	Absence of Guidelines or Standardization	e.g., ambiguity around what constitutes acceptable replication efforts and package quality.	4 (7.7%)
R6	Practical Experience with 3R Failures	e.g., direct experience in attempting to reproduce or replicate studies, often with unsuccessful outcomes.	3 (5.8%)

Takeaway 5: The vast majority of participants recognize 3R as a major issue in Software Engineering, citing insufficient resources, systemic barriers, and the absence of clear guidelines as key contributors. They also pointed to entrenched academic and research culture, such as the "publish or perish" mindset, limited openness to negative results, and lack of prioritization for replication, as critical obstacles. These factors undermine replication efforts and reinforce the perception of a persistent Replication Crisis in the field.

Practices to ensure the 3R

We asked participants if they or their research group have adopted any practices to ensure Repeatability, Reproducibility, and Replicability, including descriptions of these practices and when they were introduced. A total of 80.2% of them reported adopting practices to support the 3R principles. This contrasts sharply with the situation in biomedical research, where only 16% of the participants reported established procedures to improve reproducibility [14].

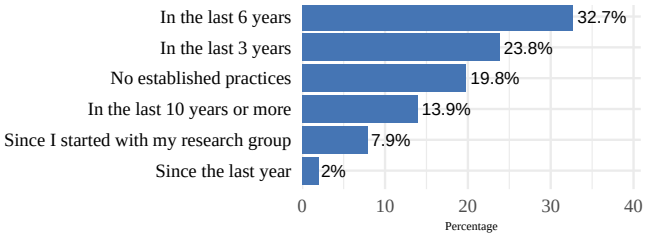

Figure 3: Time of adoption of practices to ensure Repeatability, Reproducibility, and Replicability.

As shown in Figure 3, 32.7% of the participants stated that these practices had been adopted within the past six years, suggesting growing maturity and awareness about 3R-related practices. A smaller share (2.0%) reported adopting such practices within the past year, while 13.9% indicated that they had adopted them ten years or more ago. Furthermore, 7.9% reported that these practices have been in place since they joined their research group. Almost 20% of the respondents said that they do not follow established practices to support the 3R, indicating that there is still room for broader adoption within the community.

As presented in Table 2, most of the respondents (94.7%) mentioned practices centered on ensuring that other researchers can access and reuse the data, tools, materials and procedures involved in their studies (P1). Other reported practices include clearly documenting experimental procedures, setups and threats to validity (P2); validating findings through internal checks or third party verification (P3); applying structured methods and well-established research practices (P4); aligning with Open Science initiatives, artifact evaluation processes and replication incentives (P5); and acknowledging situational or institutional constraints that can hinder replication efforts (P6). Sharing data, tools, and/or replication packages was the most cited category, cited by 72.0% of the participants. These practices demonstrate the maturity level of researchers, as well as a growing awareness and commitment within the community to improving the transparency, reliability, and cumulative value of research in Software Engineering.

Table 2: Practices to ensure the 3R.

ID	Category	Examples	Total (75)
P1	Sharing and Availability	e.g., shares data, tools, and/or replication packages, uses Docker containers or similar technologies, uses version control, and pre-registers hypotheses and analysis plans.	71 (94.7%)
P2	Documentation and Transparency	e.g., provides detailed documentation and reflects on threats to validity.	31 (41.3%)
P3	Validation and Verification	e.g., submits research for internal/external peer review, repeats experiments to confirm results, ensures independent or external validation, and conducts pre-testing or controlled testing.	22 (29.3%)
P4	Methodological Rigor	e.g., employs own framework / technique / method / process and adopts well-established scientific standards and practices.	13 (17.3%)
P5	Community and Cultural Practices	e.g., submits artifacts for evaluation, adheres to Open Science practices, and acts to promote a replication culture in the research community.	11 (14.7%)
P6	Structural and Contextual Limits	e.g., places individual responsibility on applicability of 3R practices.	1 (1.3%)

Takeaway 6: Most of the researchers reported adopting practices to support 3R, particularly sharing data and tools. These practices have become more widespread in the last six years, reflecting increased maturity and awareness. However, a non-negligible number of participants still lack formalized 3R practices, indicating room for improvement in community-wide adoption.

Frequent Contributors to 3R Failures

We asked participants to indicate their opinion on the frequency with which some factors contribute to the failure to Reproduce, Replicate, and Repeat results (Figure 4). The lack or insufficiency of protocols, computer code, or information from the original study (F6) was identified as a factor that always or very often contributes to failure in reproducing results for 75.2% of the participants.

The box plot in Figure 4 illustrates how participants rated the frequency of each factor. The most critical factors, those with higher median ratings and more responses clustered at the upper end of the scale, were the lack of availability of protocols or code (F6), raw

data not available from the original study (F2), and mistakes or lack of expertise in conducting 3R processes (F4). Selective reporting of results (F1), pressure to publish (F3), and human factors (F5) also ranked high among perceived contributors. In particular, selective reporting and publication pressure were also identified in Baker [5], Cobey et al. [14], and Murphy et al. [24] as factors that frequently contribute to irreproducible research, reinforcing their perceived relevance across disciplines. In contrast, fraud (F10), which refers to fabricated or falsified results, was considered the least frequent contributor, with lower median scores and greater response dispersion, similar to the findings of Cobey et al. [14], Baker [5], and Murphy et al. [24]. Other factors perceived as less frequent included insufficient peer review of research (F7), insufficient oversight or mentoring by the principal investigator (F9), and methods requiring technical expertise that are difficult for others to reproduce (F8).

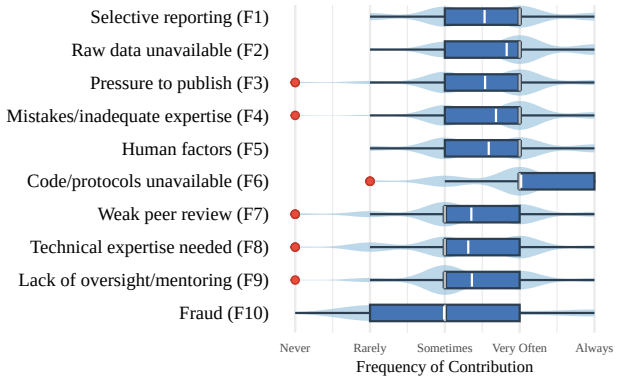

Figure 4: Frequency at which factors contribute to the failure to Reproduce, Replicate, and Repeat results.

These challenges reflect structural and procedural shortcomings in the conduct, reporting, and sharing of research. Fraud (F10) is perceived as the least influential contributor to 3R failures, which may indicate a relatively good level of trust among SE researchers and suggest that, while scientific misconduct is not dismissed, technical and systemic issues are regarded as more pervasive and pressing threats to replicability in the field. These findings suggest that failures in 3R are rarely due to a single factor but stem from a combination of cultural, technical, procedural, and other issues. Addressing these challenges requires not only better tools and standards, but also a shift in research culture toward greater transparency, collaboration, and value for replicable science.

To expand the list of factors contributing to irreproducible results, we asked participants to identify additional important contributors beyond those presented in Figure 4. Participants reported (Table 3) that irreproducibility can arise from how studies are designed, conducted, and analyzed (C2); from cultural aspects[8], lack of incentives, and systemic pressures (C3); from deficiencies in how research is communicated and documented (C4); and from barriers related to data, software, or legal and ethical constraints (C5).

The contributor most frequently cited (reported by 73.9% of the participants) was technical and infrastructure challenges (C1). These include outdated or deprecated tools used in original studies; undocumented or highly specific software dependencies and environments; incomplete, non-functional, or unavailable replication packages; and studies that are overly complex or fragile to replicate. Similarly, many of the issues raised by our participants (particularly those in categories C1 through C4) were also identified in other scientific domains, such as Biomedicine [14], indicating that these reproducibility challenges may be widespread across disciplines.

[8]Where cultural refers to prevailing norms and values within the academic community, such as the emphasis on novelty over reproducibility, limited incentives for replication, and reluctance to publish negative results.

Table 3: Contributors to irreproducible results.

ID	Category	Examples	Total (46)
C1	Technical and Infrastructure Challenges	e.g., technological obsolescence, dependency and environment issues, issues related to research artifacts, complexity of experimental setup, data variability and noise, frameworks and standardization issues, and advanced technologies.	34 (73.9%)
C2	Methodological and Statistical Issues	e.g., insufficient statistical rigor, ambiguous or imprecise definitions, sampling limitations, limited generalizability / context-dependence, human subject unpredictability, and lack of transparency in decision-making.	19 (41.3%)
C3	Cultural and Institutional Factors	e.g., cultural factors, lack of incentive for replication, publication bias and selective reporting, pressure to publish, gaps in scientific training and culture, and competitive environment.	13 (28.3%)
C4	Documentation and Communication Gaps	e.g., poor documentation or missing details, lack of standards/guidelines, misinterpretation of findings, and weak reproducibility of original findings.	11 (23.9%)
C5	Access and Availability Constraints	e.g., data unavailability or format issues, lack of access to required tools or hardware, and legal, ethical, or IP restrictions.	8 (17.4%)

Participants also noted issues such as uncontrolled variation in datasets, lack of standardized procedures or frameworks, and challenges related to the use of cutting-edge technologies. A first step toward mitigating these problems may be raising awareness of the importance of research artifacts and ensuring that they are properly documented, structured, preserved, maintained, and made available over time. Furthermore, the inherent complexity of certain subfields and the rapid evolution of technologies introduce additional reproducibility barriers that must be acknowledged and addressed. These findings highlight the need for improved artifact availability, better documentation, and more robust tools to support replicability in the face of evolving research practices and technologies.

Takeaway 7: Researchers identified the lack of protocols, raw data, and technical expertise as the most frequent causes of failures in 3R. Although misconduct like fraud was rarely cited and considered less impactful, structural and procedural issues (especially technical and infrastructure-related challenges) were considered the main threats to conducting reproducible research.

Challenges faced when attempting to replicate studies

A total of 86.1% of the respondents reported encountering barriers when trying to reproduce or replicate studies in SE. Most of these barriers (89.6%) are associated with the availability and accessibility of research artifacts (B1), including difficulties in obtaining, accessing or using essential elements such as code and tools (36.4%), data (40.3%), and other supporting artifacts (13.0%).

Table 4 shows that transparency issues related to documentation and research methods (B2) were mentioned by more than half of the participants (61.0%), which includes insufficient documentation of the experimental design, procedures, or artifacts (22.1%); vague or

ambiguous descriptions of the study design (15.6%); lack of clearly defined step-by-step replication instructions (14.3%); and inconsistencies in how methods and procedures were reported (9.1%).

Table 4: Barriers and challenges faced when trying to Reproduce or Replicate studies.

ID	Category	Examples	Total (77)
B1	Research Artifacts Availability and Accessibility	e.g., lack of access to all the artifacts or partial sharing of the artifacts, lack of access to data, and unavailable or non-functional Code, Tools, or Scripts.	69 (89.6%)
B2	Documentation and Methodological Transparency	e.g., lack of detailed documentation or protocol, incomplete or inconsistent research methods or experiment instructions, insufficient detail in original study, and difficulty understanding the method or research steps.	47 (61.0%)
B3	Environmental and Contextual Constraints	e.g., difficulty reproducing the original environment, dependency on specific tools, versions or libraries, technical complexity (difficulty setting up tools/environments), and infrastructure limitations (hardware, costs, use of cutting-edge or unstable technologies, etc.).	30 (39.0%)
B4	Contextual and Theoretical Limitations	e.g., contextual uncertainty / uncontrolled variables, replication irrelevance in qualitative studies, participant sample limitations, and lack of adequately explored methods.	14 (18.2%)
B5	Communication and Community Support	e.g., lack of response from original authors, and human and organizational factors.	8 (10.4%)
B6	General Replication Barriers	e.g., unability to replicate.	1 (1.3%)

Other reported challenges include technical barriers in reproducing the original experimental environment, such as configuring tools, managing dependencies, or maintaining compatibility with original settings (B3); issues resulting from the complexity and variability of empirical studies, particularly those involving qualitative methods or human participants, where replication may not be practical or intended (B4); and a lack of responsiveness or support from original authors or the broader research community, which hinders clarification and replication efforts (B5). One respondent also mentioned a general inability to replicate, without identifying specific causes (B6).

These findings highlight the main barriers and challenges faced by SE researchers when trying to reproduce or replicate studies. Understanding these challenges is crucial for developing strategies that mitigate them, emphasizing the importance of robust practices in artifact sharing, comprehensive documentation, and active community engagement to promote greater transparency and replicability in future research.

Takeaway 8: Most of the respondents experienced barriers when trying to Reproduce or Replicate studies, with the unavailability of code, data, and documentation cited as the main obstacles. Technical environment issues, methodological complexity, and lack of support from original authors were also noted, highlighting the importance of transparency, proper documentation, and community response.

Perceived effectiveness of initiatives to foster 3R

We asked participants to indicate how likely the following factors are to improve the Repeatability, Reproducibility, and Replicability of the research. The analysis reveals that participants generally believe that various initiatives can positively influence the Repeatability, Reproducibility, and Replicability of research in Software Engineering. As shown in Figure 5, the option *Adopting practices to increase Repeatability, Reproducibility, and Replicability in their own research* (I4) received the most favorable evaluation, with the majority of responses concentrated at the highest levels of agreement (84.2% of the participants chose the options Very likely and Likely), and a minimal number of participants expressed skepticism. This suggests that researchers recognize the impact of their own practices on 3R quality.

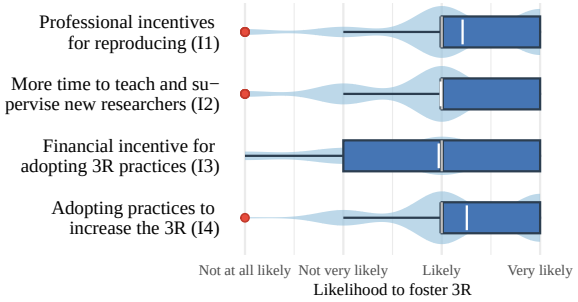

Figure 5: Factors that improve the 3R.

Professional incentives (e.g., funding or publications) to formally reproduce the work of other researchers (I1) and *Financial incentive (e.g., funding or credit toward tenure) to adopt practices that improve Repeatability, Reproducibility, and Replicability* (I3) also received high agreement, although the responses for financial incentives (I3) exhibited slightly greater variability. This indicates that while external motivations are perceived as beneficial, their effectiveness may depend on additional contextual or institutional factors.

The item *More time to teach, and better guidance/supervision, of students, postdocs, and other interns for Repeatability, Reproducibility and Replicability of research* (I2) had the lowest median and showed the most dispersed distribution among the four statements. Although many participants agreed with its potential, others were less convinced, possibly reflecting differing views on the role of mentorship and educational capacity in influencing 3R outcomes.

Overall, the results indicate a clear perception that both individual and systemic efforts, particularly those related to adopting good practices and providing structural incentives, can make a meaningful contribution to strengthening the 3R culture in SE.

Takeaway 9: Participants viewed individual and institutional initiatives as key to fostering 3R practices. Adopting good practices personally was considered the most effective action, followed by professional and financial incentives. Mentorship and supervision received mixed opinions, suggesting differing views on their role in promoting reproducible research.

Identified Factors for Enhancing Research Repeatability, Reproducibility, and Replicability

We asked participants to list any other important factors they felt we may have missed and that they believed would improve the Repeatability, Reproducibility, and Replicability of the research. The participants provided several valuable suggestions, which were grouped into seven thematic categories, presented in Table 5. Most of the responses concentrated on three main categories: Recognition and Publication Support (E1), Policies and Incentives (E2),

and Tools, Guidelines, and Infrastructure (E3), each representing approximately one third of the responses.

Table 5: Factors for enhancing research Repeatability, Reproducibility, and Replicability.

ID	Category	Examples	Total (39)
E1	Recognition and Publication Support	e.g., support for replication publications, recognition of negative results, recognition of efforts, and there are already initiatives.	13 (33.3%)
E2	Policies and Incentives	e.g., mandatory 3R requirements from editors and funding bodies, financial and institutional policy change, and incentives for researchers and reviewers.	11 (28.2%)
E3	Tools, Guidelines, and Infrastructure	e.g., use of tools and infrastructures (e.g., Docker), availability of detailed guidelines and checklists, platform or protocol standardization, and standard practices.	10 (25.6%)
E4	Cultural and Community Change	e.g., cultural change, community and venue building for replication, and ethical and professional responsibility.	7 (17.9%)
E5	Review and Evaluation Practices	e.g., peer review improvements, reviewer training and guidance, improved evaluation criteria for 3R practices, and artifact review.	6 (15.4%)
E6	Research Design and Methodology	e.g., research validation mechanisms (e.g., preregistration), contextual variation and methods, and flexibility of requirements for replications.	5 (12.8%)
E7	Training and Education	e.g., training and education at all levels and integration of 3R in research curricula.	3 (7.7%)

A total of 33.3% of participants emphasized the importance of increasing visibility, value and publication opportunities for replication/reproduction studies and negative results (E1). This category includes the recognition of researchers who adopt reproducibility practices and the acknowledgment of existing initiatives that already contribute to this goal. The second category mentioned most often involves institutional and editorial policies, as well as incentive structures, both financial and professional, that promote and reward reproducible research practices (E2). These suggestions include mandatory requirements, funding strategies, and recognition systems aligned with 3R principles. The third key area focuses on the availability and adoption of practical resources that support reproducibility (E3), such as technical tools (e.g., Docker), standardized platforms and protocols, and comprehensive guidelines and checklists to assist researchers in consistently applying 3R practices.

In addition, some participants highlighted the need for broader cultural and community changes (E4), calling for a research culture that values quality over quantity in publication, embraces replications, and fosters dedicated communities and venues for 3R research. Others pointed to the importance of improving peer review and evaluation criteria (E5), including better assessment of replication packages, structured reviewer training, and artifact evaluations to ensure rigorous and fair evaluation of 3R aspects. In this category, one important point raised by respondent [P25] concerns the need for professionalization of the peer review, considering that "Reviewers are currently unpaid and researchers around the world take on this overload of work willingly but without much time to evaluate certain attributes of research quality, such as Repeatability, Reproducibility, and Replicability of research." It is important to reflect on the "publish or perish" system currently used in the SE research community and consider cultural change.

A further set of responses advocated for foundational improvements in how studies are designed and reported (E6), suggesting the adoption of preregistration, contextual considerations in replications, and greater clarity and flexibility in defining replication. Finally, participants emphasized the importance of education and training (E7), advocating for the integration of 3R practices into academic curricula at all levels to prepare future researchers to conduct and evaluate reproducible studies.

These findings suggest that meaningful progress necessitates not only effective tools and policies, but also cultural transformation, community involvement, and long-term investment in education.

Takeaway 10: Participants emphasized the need for publication support, policy changes, and practical tools to enhance 3R. They also called for changes in research culture, improved peer review, clearer definitions and protocols, and educational efforts. These insights suggest that enhancing 3R requires a combination of structural, educational, and cultural changes in how research is conducted in Software Engineering.

RQ2.1: What are the contributing and mitigating factors of the Replication Crisis in Software Engineering?
Researchers pointed to a web of interrelated causes behind the Replication Crisis in Software Engineering. The Replication Crisis in Software Engineering stems from a combination of technical, cultural, and systemic issues. Key contributing factors include the lack of access to essential research artifacts (e.g., code, data, protocols), research culture pressures (like the emphasis on publication quantity), inadequate documentation and guidelines, and insufficient expertise for conducting replications. Fraud was perceived as a minor contributor, suggesting that the problem lies more in research practices than in misconduct.

Despite this, 80.2% of the participants reported adopting 3R practices, such as sharing replication packages, documenting procedures, and aligning with Open Science. These indicate increasing awareness and commitment within the community.

The participants also proposed strategies to improve 3R. The most cited were recognition and publication support for replication efforts, incentive and policy changes, and better tools, infrastructure, and guidelines. Others included calls for cultural change, peer review improvements, preregistration and better study design, and education and training on 3R practices.

Overall, the findings reveal that addressing the replication crisis in Software Engineering requires more than isolated efforts. It requires a systemic approach that combines technical, cultural, institutional, and educational strategies to build a more reproducible and trustworthy research environment.

5 Conclusions

Our study provides empirical insights into how Software Engineering researchers perceive the Replication Crisis and the 3R (Repeatability, Reproducibility, and Replicability) concepts. While partially aligned with the previous findings of Baker [5], our results reveal challenges specific to SE. A significant proportion (74.3%) of participants recognized a Replication Crisis in SE. Participants largely agreed with the ACM's standardized 3R definitions, particularly that of Repeatability. However, concerns arose about their clarity, relevance to qualitative research, and overlaps between terms. This highlights the need for more inclusive, precise, and context-sensitive definitions tailored to SE's methodological diversity.

SE researchers echoed broader concerns including lack of data, protocols, and pressure to publish, but also cited unique challenges such as outdated tools, dependencies, and the absence of standardized replication guidelines. Previous studies highlight Open Science's importance, but SE researchers emphasize the need for cultural shifts, peer review reforms, and greater recognition of replication studies.

Based on their experiences, participants identified contributions to the crisis and proposed a variety of solutions, from stronger publication and policy incentives to cultural changes that reward replication work. The suggestions also emphasized the need for improved peer review, better infrastructure, and training in Reproducibility practices. Addressing the Replication Crisis in SE will require a multifaceted approach that combines individual responsibility, technical improvements, cultural transformation, and systemic support.

Proposed directions. Future work must be conducted to adequately map the factors that differentiate and characterize Replicability, Reproducibility, and Repeatability, such as number of attempts, type of research (qualitative versus quantitative), what needs to be changed or kept to do each 3R, considering the original study and the peculiarities of the research to be rerun. To improve the clarity and applicability of current 3R definitions, future work could focus on qualitative research to refine and extend the terminology to better accommodate different research paradigms. To address the widespread challenges associated with 3R in SE, future efforts should aim at developing clearer guidelines, standardized protocols, and practical tools to support 3R research. This includes promoting the use of shared repositories, templates, and containerization to reduce technical barriers, and providing training and mentorship opportunities to enhance technical expertise. Institutional incentives (e.g., funding and recognition) should be aligned with 3R-friendly practices. To drive broader adoption, community-wide initiatives should also target policy and cultural research changes, improved artifact review processes, and education on reproducibility concepts, especially at the graduate level.

5.1 Threats to Validity

Internal. The questionnaire was validated by experienced researchers within our group as well as by external researchers.

To ensure that only the intended researchers responded, we sent the survey link directly to the email addresses extracted from their publications. Although we could not verify identities due to anonymity (to provide a secure environment in which participants could express their opinions), we did not observe duplicate or suspicious responses indicative of misuse or automation. Nonetheless, we acknowledge this as a potential TTV of our study.

Construct. Conducting an online survey introduces a potential threat, as it may be more difficult for participants to ask questions or seek clarification on the procedure. However, we mitigated this by providing an email address through which participants could contact us with any questions regarding the study. This threat adds some uncertainty to the results, but it is inherent of any survey, not only those conducted online. To further reduce this risk, the questionnaire was previously validated by experienced SE researchers.

We provided clear definitions of key terms such as *Replication Crisis*, *Repeatability*, *Reproducibility*, and *Replicability* to minimize variations in interpretation. Furthermore, the questionnaire was validated by experienced SE researchers to ensure that it accurately captured the participants' perceptions. However, presenting ACM-endorsed definitions beforehand may have inadvertently influenced some participants to align their responses with those definitions, rather than expressing their personal understanding or experiences.

Another threat to our study concerns the search strategy adopted to identify primary studies. We used the search string proposed by Bezerra et al. [10], which does not explicitly include the terms *Reproducibility* and *Repeatability*, originally present in the search string from Silva et al. [31]. We acknowledge that omitting these terms may have limited the scope of our review by excluding potentially relevant studies and thereby affecting the representativeness of the identified corpus. This limitation might have prevented us from reaching researchers whose work also focuses on replication but was not captured by our search strategy. To address this issue, future work will involve refining the search string to include missing terms and rerunning the search. These newly identified studies will be analyzed and compared with current results to assess potential differences and enrich our findings.

External. Although understanding the broader SE community's perspective is valuable, our study focused exclusively on authors involved in replication studies. Therefore, the findings cannot be generalized to the entire SE community, and broader generalization is left for future work. Two main factors may limit external validity: respondents' geographic concentration and the relatively low response rate (~9.5%). Nearly half of the participants are from Europe, which limits the diversity of perspectives. The low response rate also introduces potential non-response bias. One possible explanation is time-related, as some invitations may not have reached their intended recipients due to outdated email addresses or retirements.

Artifacts Availability

Although the authors' emails were publicly available in their papers, we will not make the list of collected email addresses available to prevent its use for non-academic purposes.

Some researchers provided responses that could allow their re-identification, such as mentioning their own studies as examples of reading material. As we are committed to ensuring the participants' anonymity, we opted not to consult the cited studies. The responses following this pattern were thoroughly anonymized before the data were made public available. This decision reaffirms our commitment to the privacy and security of participant data in accordance with established legal principles.

We made our scripts, data, and results publicly available at GitHub: https://github.com/IvanildoAzevedo/SurveyRepository, and permanently archived at Zenodo: DOI 10.5281/zenodo.15508756.

ACKNOWLEDGMENTS

This work was partially supported by INES 2.0, CNPq grant 465614/2014-0, FACEPE grants APQ-0399-1.03/17 and APQ/0388-1.03/14, and CAPES grant 88887.136410/2017-00. Ivanildo Azevedo is partially supported by CAPES. Sergio Soares is partially supported by the CNPq grant 306000/2022-9.

References

[1] ACM. 2020. Artifact Review and Badging - Current. https://www.acm.org/publications/policies/artifact-review-and-badging-current

[2] ACM. 2020. Artifact Review and Badging – Version 1.0 (not current). https://www.acm.org/publications/policies/artifact-review-badging

[3] Carlos E. Anchundia and Efraín R. Fonseca. 2020. Resources for Reproducibility of Experiments in Empirical Software Engineering: Topics Derived From a Secondary Study. *IEEE Access* 8 (2020). doi:10.1109/ACCESS.2020.2964587

[4] Benjamin Antunes and David R.C. Hill. 2024. Reproducibility, Replicability and Repeatability: A survey of reproducible research with a focus on high performance computing. *Computer Science Review* 53 (2024), 100655. doi:10.1016/j.cosrev.2024.100655

[5] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533 (May 2016), 452–454. doi:10.1038/533452a

[6] Maria Teresa Baldassarre, Jeffrey Carver, Oscar Dieste, and Natalia Juristo. 2014. Replication types: towards a shared taxonomy. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, England, United Kingdom) *(EASE '14)*. Association for Computing Machinery, New York, NY, USA, Article 18, 4 pages. doi:10.1145/2601248.2601299

[7] Timo Balz and Fabio Rocca. 2020. Reproducibility and Replicability in SAR Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 3834–3843. doi:10.1109/JSTARS.2020.3005912

[8] Lorena A. Barba. 2018. Terminologies for Reproducible Research. arXiv:1802.03311 [cs.DL] https://arxiv.org/abs/1802.03311

[9] C.G. Begley and L.M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483, 7391 (2012). doi:10.1038/483531a

[10] Roberta M. M. Bezerra, Fabio Q. B. da Silva, Anderson M. Santana, Cleyton V. C. Magalhaes, and Ronnie E. S. Santos. 2015. Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study. In *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 1–4. doi:10.1109/ESEM.2015.7321213

[11] Alex Borges, Waldemar Ferreira, Emanoel Barreiros, Adauto Almeida, Liliane Fonseca, Eudis Teixeira, Diogo Silva, Aline Alencar, and Sergio Soares. 2015. Support mechanisms to conduct empirical studies in software engineering: a systematic mapping study. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering* (Nanjing, China) *(EASE '15)*. Association for Computing Machinery, New York, NY, USA, Article 22, 14 pages. doi:10.1145/2745802.2745823

[12] Virginia Braun and Victoria Clarke and. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806 arXiv:https://doi.org/10.1080/2159676X.2019.1628806

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (01 2006), 77–101. doi:10.1191/1478088706qp063oa

[14] Kelly D. Cobey, Sanam Ebrahimzadeh, Matthew J. Page, Robert T. Thibault, Phi-Yen Nguyen, Farah Abu-Dalfa, and David Moher. 2024. Biomedical researchers' perspectives on the reproducibility of research. *PLOS Biology* 22, 11 (11 2024), 1–15. doi:10.1371/journal.pbio.3002870

[15] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (2020). doi:10.1145/3360311

[16] K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three Dimensions of Reproducibility in Natural Language Processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1025/

[17] Daniel Amador Dos Santos, Eduardo Santana de Almeida, and Iftekhar Ahmed. 2022. Investigating replication challenges through multiple replications of an experiment. *Information and Software Technology* 147 (2022), 106870. doi:10.1016/j.infsof.2022.106870

[18] Larissa Falcao, Waldemar Ferreira, Alex Borges, Vilmar Nepomuceno, Sergio Soares, and Maria Teresa Baldassare. 2015. An Analysis of Software Engineering Experiments Using Human Subjects. In *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 1–4. doi:10.1109/ESEM.2015.7321185

[19] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8, 341 (2016), 341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027 arXiv:https://www.science.org/doi/pdf/10.1126/scitranslmed.aaf5027

[20] Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community expectations for research artifacts and evaluation processes. In *Proceedings of the 28th ACM Joint Meeting on ESEC/FSE*. ACM. doi:10.1145/3368089.3409767

[21] Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis. *Science* 359, 6377 (2018), 725–726. doi:10.1126/science.359.6377.725 arXiv:https://www.science.org/doi/pdf/10.1126/science.359.6377.725

[22] John Ioannidis. 2005. Why Most Published Research Findings Are False. *PLoS medicine* 2 (2005). doi:10.1371/journal.pmed.0020124

[23] Cleyton V.C. de Magalhães and Fabio Q.B. da Silva. 2013. Towards a Taxonomy of Replications in Empirical Software Engineering Research: A Research Proposal. In *2013 3rd International Workshop on Replication in Empirical Software Engineering Research*. 50–55. doi:10.1109/RESER.2013.10

[24] Jennifer Murphy, Cristian Mesquida, and Joe Warne. 2023. A Survey on the Attitudes Towards and Perception of Reproducibility and Replicability in Sports and Exercise Science. *Communications in Kinesiology* 1, 5 (May 2023). doi:10.51224/cik.2023.53

[25] Brian A. Nosek and Timothy M. Errington. 2020. What is replication? *PLOS Biology* 18, 3 (03 2020), 1–8. doi:10.1371/journal.pbio.3000691

[26] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). doi:10.1126/science.aac4716

[27] Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (2011), 1226–1227. doi:10.1126/science.1213847 arXiv:https://www.science.org/doi/pdf/10.1126/science.1213847

[28] Hans E. Plesser. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics* 11 (2018). doi:10.3389/fninf.2017.00076

[29] Klaus Schmid, Sascha El-Sharkawy, and Christian Kröher. 2019. *Improving Software Engineering Research Through Experimentation Workbenches*. doi:10.1007/978-3-030-30985-5_6

[30] M. Shepperd. 2016. Replicated results are more trustworthy. In *Perspectives on Data Science for Software Engineering*, Tim Menzies, Laurie Williams, and Thomas Zimmermann (Eds.). Morgan Kaufmann, Boston. doi:10.1016/B978-0-12-804206-9.00052-0

[31] Fabio Silva, Marcos Suassuna, César França, Alicia Grubb, Tatiana Gouveia, Cleviton Monteiro, and Igor Santos. 2012. Replication of empirical studies in software engineering research: A systematic mapping study. *Empirical Software Engineering* 19 (09 2012). doi:10.1007/s10664-012-9227-7

[32] D.I.K. Sjoeberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.-K. Liborg, and A.C. Rekdal. 2005. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering* 31, 9 (2005), 733–753. doi:10.1109/TSE.2005.97

[33] Eudis Teixeira, Liliane Fonseca, and Sergio Soares. 2018. Threats to validity in controlled experiments in software engineering: what the experts say and why this is relevant. In *Proceedings of the XXXII Brazilian Symposium on Software Engineering* (Sao Carlos, Brazil) *(SBES '18)*. Association for Computing Machinery, New York, NY, USA, 52–61. doi:10.1145/3266237.3266264

[34] 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*. ACM. doi:10.1145/3313831.3376448