Reproducibility Practices of Software Engineering Controlled Experiments: Survey and Prospective Actions

André F. R. Cordeiro[a] and Edson Oliveira Jr[b]
Informatics Department, State University of Maringá, Colombo Avenue - 5790, Maringá, Brazil

Keywords: Reproducibility, Research Opportunities, Review, Software Engineering Research, Systematic Mapping, Controlled Experimentation.

Abstract: Reproducibility can be described as a characteristic that contributes to expanding knowledge in science. This paper investigates the reproducibility of experiments in Software Engineering (SE) in a context where the literature points to challenges in verifying experimental results. The central problem addressed is the difficulty in reproducing experiments in SE due to the different factors, such as sharing and artifact management. We then aimed to identify the factors necessary to achieve reproducibility in SE experiments, characterizing these factors in terms of the reproducibility crisis, experimental workflows, research practices, FAIR principles application, and reproducibility improvements. We planned and conducted one survey with 16 participants who answered a questionnaire with 33 questions. The results show that most participants perceive a reproducibility crisis in the field and point to factors such as lack of public data and incomplete information on methods and experimental setups as the main causes. Furthermore, the results highlight the importance of sharing data, metadata, and information about research teams. We also provide points to possible actions to improve reproducibility in SE experiments. The contributions include a detailed analysis of the challenges to reproducibility in SE, as well as the identification of practices and measures that can improve reproducibility.

1 INTRODUCTION

The literature contains different studies describing experimental activities in software engineering (SE). An assessment of methods for verifying experimental findings is presented in Juristo and Vegas (2010). The results indicate that reanalysis, replication, and reproduction are frequently considered in SE research.

Reproducibility, often referred to as reproduction, is a fundamental principle for advancing knowledge across scientific fields Juristo and Vegas (2010). It is achieved when identical results are obtained using the same methodology, even when experiments are conducted in different laboratories, by different operators, and with varying equipment Kitchenham et al. (2020).

Controlled experiments in SE often face challenges related to reproducibility. These challenges encompass aspects such as availability, standardization, review processes, and the generation and evolution of experimental artifacts (Solari et al., 2018). To address these challenges, this paper surveys SE researchers who have experienced experimentation. The survey aims to deepen the understanding of reproducibility issues in the field.

The structure of this study is as follows. Section 2 discusses the background and related work. Section 3 outlines the research methodology. Section 4 presents the findings, followed by a discussion in Section 5. Section 6 discusses the validity evaluation of this study. Prospective actions are proposed in Section 7, and final remarks are provided in Section 8.

2 BACKGROUND AND RELATED WORK

This section outlines the theoretical foundation for the study, organized into controlled experiments in software engineering, reproducibility in software engineering, and related research.

[a] https://orcid.org/0000-0001-6719-2826
[b] https://orcid.org/0000-0002-4760-1626

Cordeiro, A. F. R. and Oliveira Jr, E.
Reproducibility Practices of Software Engineering Controlled Experiments: Survey and Prospective Actions.
DOI: 10.5220/0013475600003929
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 2*, pages 372-379
ISBN: 978-989-758-749-8; ISSN: 2184-4992
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

2.1 Controlled Experimentation in Software Engineering

Experimental Software Engineering (ESE) is a research area focused on investigating best practices for conducting experiments in SE (Wohlin et al., 2012). A controlled experiment in SE can be described as a process that involves several phases, including definition, planning, and execution (Wohlin et al., 2012). Figure 1 illustrates the experimental process.

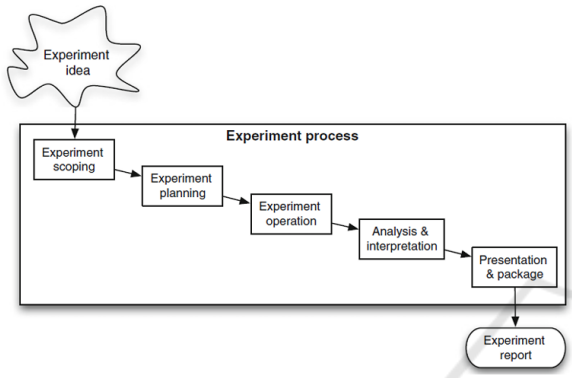

Figure 1: Experimental process (Wohlin et al. (2012)).

During the definition phase, the scope of the experiment is established, and specific objectives are determined. The planning phase involves organizing the experiment by selecting the context, formulating hypotheses, identifying variables, choosing participants (if required), deciding on an experimental design, preparing instrumentation, and assessing validity.

The operation phase is where the experiment is executed according to the protocol defined during planning. Afterward, the collected data is analyzed and interpreted to determine whether the hypothesis can be rejected. All experimental artifacts, including data, should be carefully packaged for presentation, dissemination, and potential future reproduction. Finally, detailed experimental reports are prepared to document findings and methodologies (Wohlin et al., 2012).

2.2 Reproducibility in Software Engineering

Reproducibility in SE can be examined across various levels, contexts, or subareas (Li, 2021; Kitchenham et al., 2020). The study of Liu et al. (2021) investigates the urgency and importance of reproducibility and replicability in the application of Deep Learning (DL) in SE.

Despite the availability of various solutions, including processes and tools, the reproducibility problem persists, particularly in the context of empirical studies (Li, 2021; Kitchenham et al., 2020). This challenge is especially perceived in controlled experiments (Liu et al., 2021; Anchundia et al., 2020).

2.3 Related Work

Kitchenham et al. (2020) provides an assessment of the reproducibility and validity of experimental results in SE. The study involved a systematic review of research published in SE journals. Similarly, Li (2021) investigates issues in the Evidence-Based Software Engineering (EBSE) subarea, focusing on challenges such as the reuse of search strings and the lack of reproducibility in automatic searches. From these studies, it was possible to observe evidence of the reproducibility problem in SE, at least in terms of experimental results and secondary studies.

An analysis of tools to enhance reproducibility in SE experiments is presented in Anchundia et al. (2020). It also explores the role of community acceptance in adopting tools and practices. This study influenced the understanding of possible influences of research tools and practices to achieve reproducibility.

To better understand reproducibility challenges across scientific fields, Samuel and König-Ries (2021) conducted a survey focusing on planning and execution practices. The study examined the concept of a reproducibility crisis, the application of FAIR principles, and measures to ensure reproducibility. Key areas covered include Chemistry, Biology, and Computer Science. The findings of Samuel and König-Ries (2021) highlight that reproducibility issues are not limited to SE but represent a widespread problem across multiple disciplines.

The perceptions resulting from the aforementioned studies influenced the definition of the scope of this study, within the SE.

3 RESEARCH METHODOLOGY

The survey described in this paper considers a methodology based on goals, research questions, target audience, population, sampling, instrument and evaluation, and data sharing. We structured the survey according to the guidelines specific to SE, presented in the literature (Linåker et al., 2015).

3.1 Goal and Research Questions

This study **aims to** identifying factors needed to achieve SE experiment results' reproducibility, **with**

the purpose of characterizing such factors, **with respect to** reproducibility crisis, experiment workflows and research practices, measures to ensure of reproducibility of results, the introduction of FAIR data principles, and research practices for improving reproducibility, **from the perspective of** SE researchers and practitioners, **in the context of** the SE research community.

Therefore, the main research question that guided this study was: '**How is reproducibility understood and applied in software engineering?**'. To answer this question, we define the following secondary research questions (SRQ):

- **SRQ1.** What leads to a reproducibility crisis in software engineering?
- **SRQ2.** What are the different experiment workflows and research practices followed in software engineering?
- **SRQ3.** What are the current measures taken in software engineering to ensure the reproducibility of results?
- **SRQ4.** Has the introduction of FAIR data principles influenced the research practices in software engineering?
- **SRQ5.** Which research practices could improve reproducibility in software engineering?

SRQ1 assumes the existence of a reproducibility crisis in SE, as suggested by studies such as Li (2021) and Kitchenham et al. (2020). This assumption is based on the observed lack of reproducibility efforts and a limited number of reproductions in SE studies. To explore this, researchers are asked about their perspectives on this potential crisis and the factors that may contribute to reduced reproducibility.

SRQ2 investigates the workflows and practices employed in SE research. These workflows involve the management of artifacts, data, and storage devices, which play a crucial role in scientific research processes.

SRQ3 posits that specific measures can help ensure the reproducibility of experimental results in SE. Examples include the ease of retrieving experimental data and metadata, the ability to reproduce published results, and the effective execution of reproductions. These factors are considered critical in assessing the assurance of reproducibility.

SRQ4 explores the role of the FAIR Data Principles in promoting reproducibility. It seeks to determine whether these principles are known and applied by SE researchers. The assumption is that familiarity with and application of the FAIR principles significantly influences research practices in SE.

SRQ5 focuses on identifying research practices that can improve reproducibility in SE. It assumes that sharing data, metadata, and information about research teams can contribute to enhanced reproducibility.

3.2 Target Audience and Population

This study considered researchers and practitioners in the SE area who are knowledgeable about experimentation.

3.3 Sampling

This session presents the profile of the 16 participants in this study. Details on the areas of activity and study in SE are presented in Figure2. Most of the participants (11) work as university professors. In addition to these professionals, the participation of students (undergraduate and graduate) and researchers (participant and leader of the research group) was also observed. A professional who works as a developer also participated.

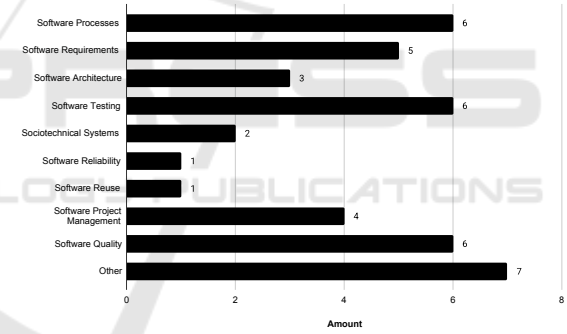

Figure 2: Software Engineering Participants' Areas.

Figure 2 presents the SE areas considered by the participants. Each participant was allowed to register one or more areas of study. When observing the results of such a figure, some areas stand out, such as Software Processes, Software Testing, and Software Quality. Practitioners who investigate other areas not mentioned as options in the form also participated. These professionals investigate problems addressed in the areas of SE Experimentation (1), SE Education (2), Human-Computer Interaction in SE context (1), Search-Based SE (1), Human Aspects of SE (1), and Software Accessibility.

3.4 Instrument and Evaluation

We adapted one questionnaire for this survey, with **33 questions** related to the **reproducibility of experi-**

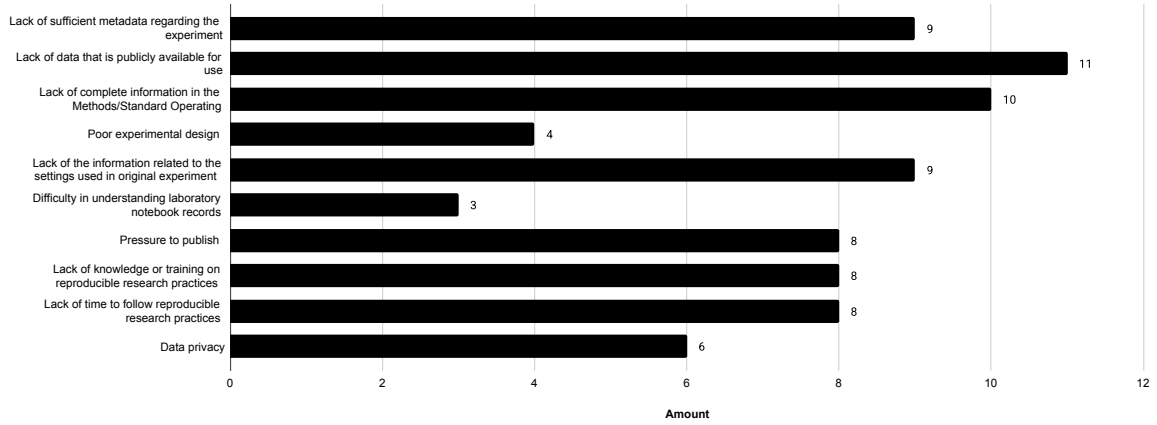

Figure 3: Factors related to poor reproducibility.

ments in SE. The base questionnaire, considered in the adaptation, was used in the study of Samuel and König-Ries (2021). We built this instrument with Google Forms[1]. Initially, we contacted the participants and after the concordance with the participation, we sent participants the access link by email.

During the development of the instrument, we estimate the response time of the questions. We evaluated the instrument with a pilot project with three researchers who were not in the sampling. After the pilot project, no changes were suggested. Thus, we kept the instrument as is before the pilot project.

We presented the questions of the instrument in different formats, such as multiple choice, selection box, open box, and checkbox grid. To facilitate interpretation, we prepared optional questions with short statements.

3.5 Study Data Availability

Data from this paper is available at https://zenodo.org/records/14888972.

4 RESULTS

This section presents the results of our survey regarding the five research questions from Section 3.1.

4.1 SRQ1 - Reproducibility Crisis

The reproducibility crisis refers to the growing belief that the results of many scientific studies are difficult or impossible to reproduce after further investigation, either by independent researchers or by the original researchers themselves (Kitchenham et al., 2020).

[1]https://www.google.com/intl/pt-BR/forms/about/

Fourteen participants registered a perception about the existence of this problem. These results represent a previous perception observed in other studies as Li (2021) and Kitchenham et al. (2020). Considering the existence of a reproducibility difficulty, Figure 3 presents possible factors associated with such difficulties.

Among the factors that may explain low reproducibility and that were considered in this study, we can mention the "**lack of data that is publicly available for use**", registered by eleven participants.

4.2 SRQ2 - Experiment Workflows and Research Practices

Experiment workflows and research practices can be understood as guidelines and tools for conducting research in SE.

Considering the locations used for storing experimental data and metadata, Figures 4 and 5 present the details, respectively. Figure 4 presents that some of

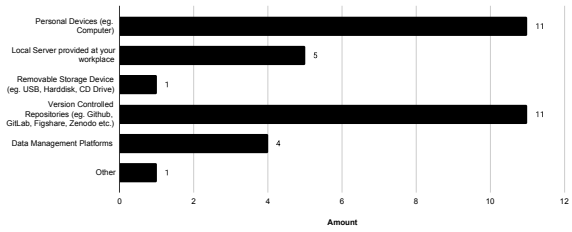

Figure 4: Local of storage of the experimental data files.

the experimental data storage locations cited include "**personal devices**" and "**version controlled repositories**." Regarding metadata, Figure 5 presents the main current storage options, such as "**handwritten lab notebooks**", "**electronic notebooks**", and "**data management platforms**". Six participants recorded that they write scripts and programs. Four participants

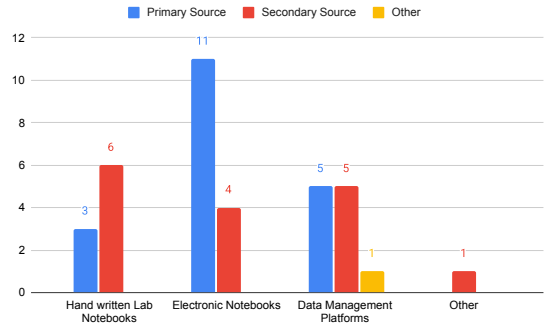

Figure 5: Local for storage of the experimental metadata.

recorded that they do not write scripts.

In addition to the workflows and practices presented in Figures 4, and 5, participants also reported the importance of detailed documentation on the infrastructure used in the experiment, with the objectives of maintaining traceability and facilitating reproducibility.

4.3 SRQ3 - Measures to Ensure Reproducibility of Results

Initially, there is a supposition that different actions can be taken to ensure reproducibility. Figures 6 and 7 present details related to possible actions.

Figure 6 presents different data sets to investigate the ease of retrieval of experimental data in the context of a researcher participating in the planning and execution of an experiment. Considering the context

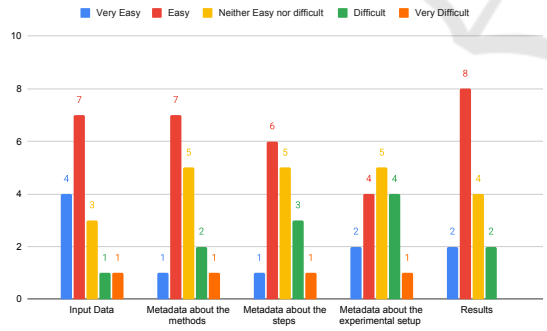

Figure 6: Facility to find all the experimental data.

of a new participant in a research team without any instruction, Figure 7 presents the results of the evaluation of the ease of retrieval of experimental data. Regarding the reproduction of published results from others, it was found that eight participants reported that they "**never tried to reproduce other published results**". Despite possible difficulties with the reproduction of published studies, most participants did not experience problems related to the reproduction of their studies. Only two participants reported having

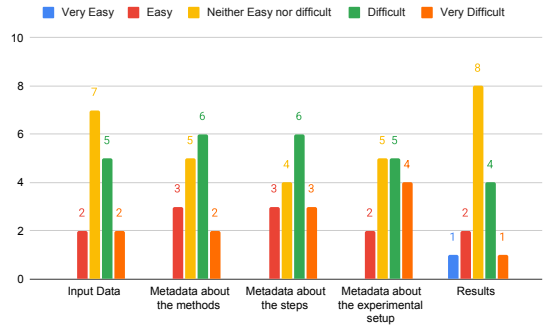

Figure 7: Facility for a newcomer member to obtain all the experimental data without any instructions.

been contacted about reproducing published studies.

Considering the reproduction of their own experiments to verify the results, eight participants reported that they **sometimes perform this reproduction to verify**. Five participants also reported that they **do not consider this activity**.

4.4 SRQ4 - FAIR Data Principles

The FAIR principles can be applied to artifacts in general. Data sets represent examples. The acronym FAIR stands for the combination of the characteristics **Findable**, **Accessible**, **Interoperable**, and **Reusable**. Twelve participants are aware of the FAIR principles. Although this knowledge is positive, it was also observed that six participants **heard about these principles but did not know what they meant**. Regarding the application of the principles, Figure 8 presents the details.

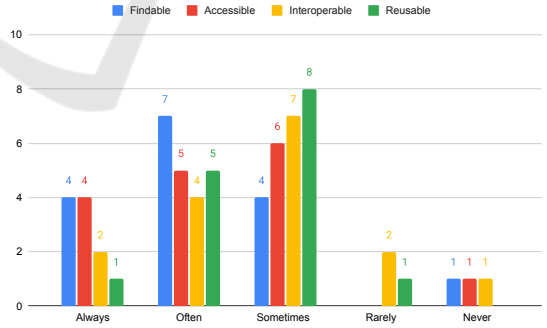

Figure 8: Application of the FAIR principles in the research.

Considering the principles findable, accessible, interoperable, and reusable, it can be seen that **some principles are considered more than others in terms of frequency**.

4.5 SRQ5 - Research Practices to Improve Reproducibility

Initially, there is a premise that different factors are important to understanding a scientific experiment in SE, with the purpose of enabling reproducibility. Figures 9, 10, 11, and 12present details of the sharing of data and metadata, and the knowledge of the research team involved as potential factors that contribute to reproducibility.

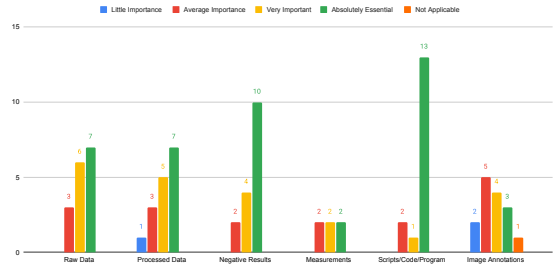

Figure 9: Sharing of experimental data.

Regarding the sharing of experimental data, Figure 9 shows that participants classified as "**absolutely essential**" the sharing of "**raw data**", "**processed data**", "**negative results**", "**measurements**", "**script/code/program**".

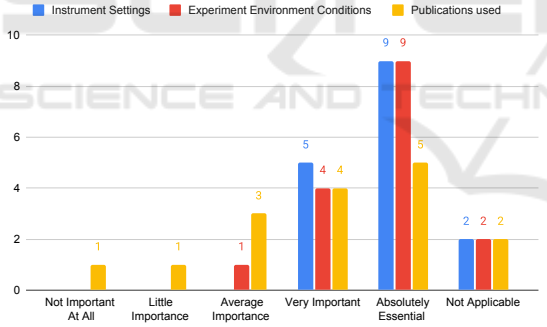

Figure 10: Sharing of metadata regarding settings.

Figure 10 presents the evaluations on the sharing of metadata related to experiment settings. The majority of participants also evaluated as "**absolutely essential**" the sharing of "**instruments settings**", "**experiment environment conditions**", and "**publications used**".

Figure 11 presents results on the sharing of metadata related to steps and plans related to the experiment. The sharing metadata was considered "**absolutely essential**", in terms of "**methods**", "**activities/steps**", "**order of activities/steps**", "**validation methods**" and "**quality control methods**".

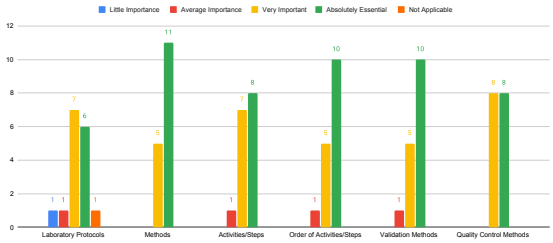

Figure 11: Sharing of metadata regarding all the steps and plans.

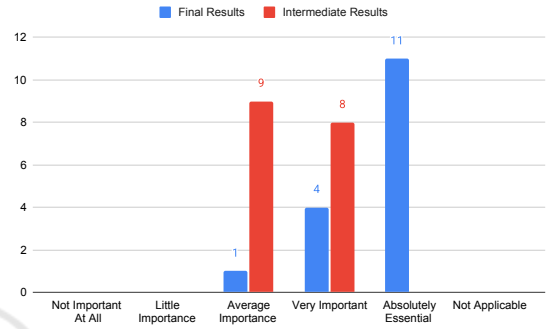

Figure 12: Sharing of the intermediate and final results of each step of the experiment.

Regarding the sharing of experiment results, Figure 12 presented different evaluation results. The "**final results**" were evaluated as "**absolutely essential**" and the "**intermediate results**" were evaluated as "**average importance**".

In addition to sharing experimental data and metadata and knowledge about the research team (Figures 9, 10, 11, 12), participants also reported the importance of sharing detailed descriptions of experiment limitations, as well as justifications for methodological choices, previous versions of scripts and data for traceability, cross-validation reports, and third-party reanalysis.

5 DISCUSSION OF RESULTS

This section discusses the results presented in Section 4.

5.1 Reproducibility Crisis

Considering the results presented in section 4, an apparent reproducibility crisis in SE is observed. Different factors may explain the difficulty of reproducing an experiment. Among them are the **lack of publicly available data,** the lack of complete information on the methods employed, **lack of information on the settings** used in the experiment.

5.2 Experiment Workflows and Research Practices

Different **kinds of artifacts** can be considered in experiments, such as scripts, diagrams, and code. Different **types of data** are also considered. Tabulars, measurements, metrics, and graphs are examples. As for data storage locations, personal devices, version control repositories, and institutional repositories were mentioned.

5.3 Measures to Ensure Reproducibility of Results

In general, the researchers responsible for the experiment considered it **easy to recover experimental data** for both input data and results, as well as the metadata about the methods, steps, and experimental setup. In the case of data retrieval to be performed by a novice researcher without any instruction, it was observed that it was **difficult to obtain metadata** about methods, steps, and experimental setup.

Regarding the reproduction of experiments to verify results, we observed that the majority of participants reported that they did not perform this activity, which is a concerning factor.

5.4 FAIR Data Principles

Considering the knowledge about the FAIR principles, we find that a considerable number of participants **know the principles**, even if they do not **know exactly what they mean**. About application, we acknowledge that some principles can be **applied more frequently**. This is the case of the Findable and Accessible principles, which are essential characteristics for effective research data.

5.5 Research Practices to Improve Reproducibility

The **sharing of experimental data** is considered absolutely essential by participants. Regarding the **sharing of metadata** about materials, and settings, time, duration, location, steps, and software used in the experiment, the participants also considered it absolutely essential. For the **sharing of results**, we saw that the **sharing of intermediate results** was evaluated as medium importance. The sharing of final results is considered absolutely essential.

6 VALIDITY EVALUATION

In this section, we discuss the main threats to the validity of our survey based on the guidelines by Linåker et al. (2015) regarding face, content, criterion, and construct. To ensure **face validity**, the survey form was reviewed by three researchers during a pilot project. We focused on achieving **content validity** by conducting an unstructured interview with researchers experienced in experimentation in SE to review the questionnaire.

To address **criterion validity**, we organized the questionnaire into distinct sections, each corresponding to a specific research question. To assess **construct validity**, we conducted activities during the instrument pilot project, the interviews, and the literature review.

7 PROSPECTIVE ACTIONS

The observed results and respective discussions present clear evidence that reproducibility must be addressed in prospective investigations. Therefore, we will provide some actions to be taken regarding the topic discussed.

Data and metadata storage options should be analyzed regarding different options, assessing their advantages and disadvantages. It might include personal devices, version-controlled repositories, local servers, data management platforms, and electronic or physical lab notebooks. This analysis can aid in **identifying best practices for ensuring data accessibility and integrity, facilitating the reproducibility of experiments**.

Creating frameworks for sharing data and metadata that can include data repositories, tools for metadata management, and best practice guidelines might **encourage data sharing**. These frameworks can help to overcome the lack of availability of data and complete information on methods and settings, which have been identified as the main problems for reproducibility.

8 FINAL REMARKS

Based on the results presented and discussed in the paper, it is clear that there is a need to address the issue of reproducibility in future research in the field of SE. This study corroborated an apparent reproducibility crisis in the field, caused mainly by the lack of public data and incomplete information about the

methods used and the experiment settings. This crisis not only makes it difficult to verify research results.

We also identified several factors that influence reproducibility in SE experiments, such as the different types of artifacts used , the types of data considered , and the storage locations of these data and metadata. Our survey emphasized the importance of sharing raw and processed data, negative results, measurements, scripts/code/programs, as well as metadata related to materials, and settings.

We acknowledge essential points to be investigated for future actions, such as analysis of the advantages and disadvantages of the various forms of data and metadata storage; and development of frameworks to facilitate the sharing of data and metadata and promote reproducibility in SE experiments.

ACKNOWLEDGMENTS

The authors thank the participants for their collaboration with the research. Edson OliveiraJr thanks CNPq/Brazil Grant #311503/2022-5.

REFERENCES

Anchundia, C. E. et al. (2020). Resources for reproducibility of experiments in empirical software engineering: Topics derived from a secondary study. *IEEE Access*, pages 8992–9004.

Juristo, N. and Vegas, S. (2010). Replication, reproduction and re-analysis: Three ways for verifying experimental findings. In *RESER*.

Kitchenham, B., Madeyski, L., and Brereton, P. (2020). Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment. *EMSE*, 25:353–401.

Li, Z. (2021). Stop building castles on a swamp! the crisis of reproducing automatic search in evidence-based software engineering. In *ICSE-NIER*, pages 16–20. IEEE.

Linåker, J., Sulaman, S. M., Maiani de Mello, R., and Höst, M. (2015). Guidelines for conducting surveys in software engineering.

Liu, C., Gao, C., Xia, X., Lo, D., Grundy, J., and Yang, X. (2021). On the reproducibility and replicability of deep learning in software engineering. *TOSEM*, 31(1):1–46.

Samuel, S. and König-Ries, B. (2021). Understanding experiments and research practices for reproducibility: an exploratory study. *PeerJ*, 9:e11140.

Solari, M., Vegas, S., and Juristo, N. (2018). Content and structure of laboratory packages for software engineering experiments. *IST*, 97:64–79.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS