Reimagining Studies' Replication: A Validity-Driven Analysis of Threats in Empirical Software Engineering

Ivanildo Azevedo
Centro de Informática
Universidade Federal de Pernambuco
Recife, PE, Brazil
vando.aze@gmail.com

Ana Paula Vasconcelos
Centro de Informática
Universidade Federal de Pernambuco
Universidade do Estado de Mato Grosso
Recife/Barra do Bugres, PE/MT, Brazil
aninha.lfv@gmail.com

Eudis Teixeira
Coordenação de Informática
Instituto Federal do Sertão Pernambucano
Petrolina, PE, Brazil
eudis.oliveira@ifsertao-pe.edu.br

Sergio Soares
Centro de Informática
Universidade Federal de Pernambuco
Recife, PE, Brazil
scbs@cin.ufpe.br

ABSTRACT

Context: Replication studies play an important role in strengthening the empirical foundations of Software Engineering (SE). However, the existing literature reveals that the reporting of Threats to Validity (TTVs) remains inconsistent or superficial, potentially undermining the reliability of the replication results. **Objective:** The goal of this study is to analyze how replication studies consider TTVs present in original studies in SE. **Method:** We conducted a Systematic Literature Review (SLR) that resulted in 83 replication studies published between 2022 and 2024. We analyzed the presence and specificity of TTVs in four validity dimensions (construct, internal, external, and conclusion), considering different research methods and types of replication. **Results:** Our analysis shows that replication studies in Empirical Software Engineering (ESE) tend to report threats to validity more frequently and in greater detail than original studies, particularly with regard to external and internal validity. Nevertheless, threats related to the validity of the conclusion and construct remain underreported. We observed that controlled experiments generally address the different types of TTVs more comprehensively, whereas surveys and case studies provide more limited coverage. With respect to types of replication, close and differentiated replications are predominant, while conceptual and internal replications remain underexplored in the field. **Conclusion:** Although there is growing attention to the identification of TTVs in replication studies, reporting remains uneven across validity dimensions and study types. More structured and diverse replication strategies are needed, along with better guidelines to support comprehensive TTV reporting and enhance the rigor and methodological value of replication efforts in ESE.

KEYWORDS

Replication, Software Engineering, Open Science, TTVs, SLR.

1 Introduction

Replication plays an important role in advancing scientific knowledge by enabling the verification, validation, and extension of findings. Replication contributes to research transparency and reliability. However, despite its recognized importance, replication is often undermined by the underreporting of TTVs, which compromises both the replicability and interpretability of studies [22]. Although some initiatives have promoted the sharing of research artifacts such as data and code [29], systematic investigations into how TTVs are reported remain limited.

Despite advances, replications often suffer from incomplete documentation, making them difficult to reproduce and compromising the credibility of results [16, 32]. One of the factors limiting successful replication is the inadequate reporting of TTVs, which compromises both the reliability and generalizability of findings. Prior studies [6, 9] have emphasized the importance of assessing the reliability of empirical studies replicating results in different contexts and systematically identifying and mitigating TTVs, yet their explicit consideration in replication studies remains limited.

Recent studies highlight the limited and often superficial attention given to TTVs. About 80% of papers from Computer Security and SE conferences do not include a dedicated section to address TTVs [28]. In contrast, in papers that received the ACM SIGSOFT Distinguished Paper Award at ICSE just the minority (33%) do not present a dedicated section discussing TTVs [22]. Lago et al. [22] reported that "A large number of studies do not report a TTV section, further corroborating the trend of shallow or completely missing importance given to TTVs in empirical studies". This reinforces that careful reporting of TTVs is more common in top-ranked studies but remains inconsistent across the broader community.

Additionally, previous research has shown that differences in contextual variables can lead to divergent replication results, emphasizing the need to clearly identify and communicate TTVs [27]. However, prior mapping and review efforts on replication [7, 12, 31] have not addressed the role of TTVs and their mitigation strategies as a central aspect of fostering study replication.

This research addresses this gap by investigating how TTVs are reported, categorized, and compared between replication studies and their original counterparts in SE. Specifically, we analyze whether replication studies identify, discuss, and attempt to mitigate TTVs differently from the original studies, and how research methods and replication types influence this reporting.

The main contributions of this study are:

- **Mapping of current practices in reporting TTVs.** We examine the frequency and positioning of TTVs in both original and replication studies, identifying patterns that may inform clearer and more consistent reporting practices.
- **Influence of research methods in reporting TTVs.** We analyze how different empirical methods (e.g., experiments, case studies) relate to the types of TTVs reported, providing insights for method-sensitive replication planning.
- **Open Science.** All data and scripts of this study are publicly available, promoting transparency, reuse, and enabling future investigations. Link in: ARTIFACT AVAILABILITY.

Ultimately, this work contributes to strengthening replication practices in Software Engineering by shedding light on how TTVs are treated in replication efforts and by encouraging greater methodological rigor, clarity, and transparency in empirical research.

Paper structure. The remainder of this paper is organized as follows: Section 2 presents the key concepts and related works, Section 3 describes the methodology, Section 4 presents and discusses our findings, Section 5 concludes and outlines planned next steps. Finally, Section 5 provide a link to our research artifacts.

2 Background

Replication plays a central role in improving the reliability of empirical research. However, its effectiveness depends on how rigorously studies address threats to validity and how clearly replication strategies are defined. To support our analysis, we rely on three conceptual frameworks that provide the basis for categorizing replication types, threats to validity, and empirical methods.

2.1 Key Concepts

Baldassarre et al. [4] propose a taxonomy that classifies replications into five types: *internal* (conducted by the same research group), *external* (by different groups), *close* (with minimal changes to the original study), *differentiated* (with controlled variations), and *conceptual* (using different methods to test the same hypothesis). It allows distinguishing between replications that confirm results and those that explore generalizability under new conditions. It will enable a more detailed analysis of how TTVs manifest across different replication strategies and which approaches are more prevalent.

Wohlin et al. [36] classified TTVs in four categories that refer to different aspects of research rigor, ranging from the existence of an effect (*conclusion*), to causal inference (*internal*), alignment between theory and measurement (*construct*), and generalizability (*external*). This framework provides a systematic basis for assessing and mitigating potential risks in empirical studies.

We also use the taxonomy of empirical methods proposed by Easterbrook et al. [14], which includes *controlled experiments*, *case studies*, *surveys*, *ethnographies*, and *action research*. These methods differ in goals and approaches: experiments test causality in controlled settings; case studies examine phenomena in real-world contexts; surveys collect generalizable data; ethnographies offer deep cultural insights; and action research integrates practice with inquiry. This classification helps analyze how methodological choices influence the identification and treatment of TTVs.

Prior studies [3, 11, 19, 21, 24, 26] have highlighted a lack of standardization and clarity in the use of the 3R (Replication, Reproduction, and Repetition) across disciplines, including Software Engineering. Although some initiatives have aimed to align terminology with broader scientific communities [1], inconsistencies in definitions persist, often leading to divergent interpretations, because researchers may use the same terms to refer to different concepts, or different terms to describe the same practice. For this reason and to facilitate readability, we clarify that throughout this work, we make no distinction between these terms. When referring to replications, we broadly encompass all 3R, thus also referring to reproductions and repetitions of previous studies.

These conceptual frameworks provide the theoretical foundation for both the design of our study and the analysis of how replication practices and threats to validity are reported in the SE literature.

2.2 Related Works

2.2.1 TTVs in ESE: Gaps, Categorization, and Guidelines. The literature has extensively discussed challenges in how threats to validity are recognized and reported in Empirical Software Engineering.

Lago et al. [22] examined 91 awarded papers to understand how TTVs are addressed in the literature. Their findings revealed that discussions on these threats are rare and superficial, although 67% of studies have a TTV section. Additionally, 61.5% of the papers did not present mitigation strategies. Similarly, Verdecchia et al. [35] and Gren [20] highlight the superficial and often post-hoc treatment of TTVs, calling for earlier and more reflective integration of TTV discussions throughout the research process.

To improve standardization, several authors have proposed taxonomies and checklists. Ampatzoglou et al. [2] identified inconsistencies in how secondary studies report TTVs and proposed a classification framework. Studies such as Malhotra and Khanna [23] and Barros and Neto [5] presented tailored TTVs lists and guidelines for prediction models and search-based Software Engineering, respectively. Tools such as ValiDEPlan [13] and PrioriTTV [33] were also introduced to support the planning and prioritization of TTVs in experimental contexts.

While these contributions aim to improve reporting practices, others reveal a disconnect between researchers' awareness of TTVs and their application in practice [30, 34]. This highlights the ongoing need for clear protocols and cultural shifts in empirical Software Engineering to support better TTV handling.

Efforts to structure and document TTVs have led to proposals such as the Evidence Tetris model [37], which synthesizes evidence to support experimental design decisions, and the creation of a domain-specific Knowledge Base for TTVs [8]. These initiatives aim to promote semantic understanding and the reuse of mitigation strategies. Additionally, previous studies have connected TTV reporting to researchers' philosophical stances [25], and found inconsistent terminology and underreporting of validity threats in Software Engineering papers [18].

2.2.2 Our Contribution. Despite these initiatives, we did not meet any studies focus on how TTVs are handled specifically in the context of replication. Our study addresses this gap by analyzing how replication studies identify, categorize, and reflect on TTVs. By comparing practices across research methods. We provide a

detailed overview of how TTVs affect the reliability and generalizability of replication efforts. Furthermore, our findings support the development of structured, context-aware reporting guidelines to strengthen replication as a foundation for Empirical Software Engineering.

3 Study Design

The goal of this study is to analyze how replication studies consider TTVs present in original studies in Software Engineering. We want to understand how these threats are identified, discussed, and mitigated throughout the replication process, assessing their impact on the reliability and generalization of the results. Additionally, the study investigates the strategies adopted to address these threats, contributing to the improvement of replication practices. To reach our goal, we intend to answer the following research questions:

• **RQ1** *What is the current state of TTV reporting in both replication studies and the original studies they replicate?*

• **RQ2** *What is the relationship between research methods and the types of TTVs reported in replication studies?*

3.1 Search Strategy

We conducted a Systematic Literature Review (SLR) using an automated search in the SCOPUS engine, targeting peer-reviewed papers published in the last three years (2022, 2023, and 2024) that report replication, repetition, or reproduction of empirical research in Software Engineering (see Figure 1).

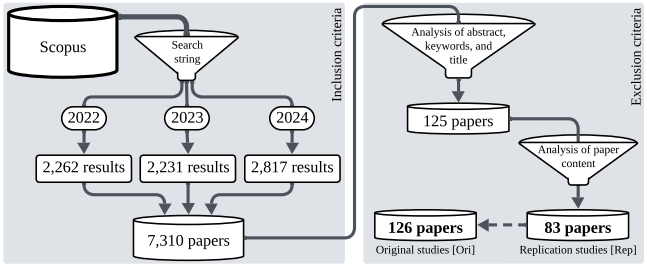

Figure 1: Study Design.

3.2 Methodological Decisions

Search Engine. We performed an automatic search using the SCOPUS engine, which is recognized as the largest database of abstracts and references in the scientific literature, covering a wide range of academic disciplines and providing access to a large number of indexed publications [15]. Cruz et al. [12] found that the leading journals and conferences were indexed in SCOPUS, indicating that the platform includes a selection of relevant studies publications in the Software Engineering research community [17].

Coverage. We selected primary studies published between 2022 and 2024 to focus our analysis on the most recent efforts related to replication and the treatment of TTVs. This allows us to investigate trends, practices, and concerns, in light of the growing attention to research transparency and reproducibility in recent years.

Search String. The search string (Figure 2) was constructed by including keywords associated with 3R (Replication, Reproduction, and Repetition), terms related to the types of empirical studies discussed in Easterbrook et al. [14], and the phrase "software engineering" to appropriately narrow the scope of the retrieved results.

Search Procedure. In the initial stage of the search, we applied the inclusion criteria (see Section 3.2.1) using the SCOPUS search engine, which returned a total of 7,310 papers. After screening titles, abstracts, and keywords, according to the exclusion criteria (presented in Section 3.2.1), 125 papers remained. In the final stage, after a full-text analysis using the exclusion criteria, a total of 83 papers were selected to be analyzed and data extracted to answer our research questions. These 83 replication [Rep] papers replicated a total of 126 original [Ori] studies[1].

(reproducibility OR reproduction OR reproductions OR reproduce OR reproduced OR reproducing OR replicability OR replication OR replications OR replicate OR replicated OR replicating OR repeatability OR repetition OR repetitions OR repeat OR repeated OR repeating)
AND
(survey OR surveys OR experimental OR experiment OR experiments OR "quasi-experiment" OR "case study" OR "case studies" OR ethnography OR "ethnographic study" OR "ethnographic studies" OR "action research")
AND
("software engineering")

Figure 2: Search String

3.2.1 Inclusion and Exclusion Criteria. To be selected for this research, the paper must meet all the following *Inclusion Criteria*:

- Peer-reviewed papers (white papers).
- Empirical studies.
- Papers published between 2022 and 2024.
- Papers that conducted replication, repetition, or reproduction of empirical studies in Software Engineering.

Based on the manual analysis of title, keywords, and abstract, studies that meet any of the following *Exclusion Criteria* were excluded:

- Not Software Engineering papers.
- Papers not written in English.
- Papers not available online.
- Duplicate publications.
- Meta-analyses, secondary, or tertiary studies.
- Papers without a specific section/subsection/highlight on TTVs.
- Grey literature.
- Studies in which the main goal of replication is to evaluate the replicability of a specific approach, including tools, methodologies, or other techniques.
- Replication studies for which the original study is not available.

3.2.2 Data Extraction. We used Google Sheets to systematically store and organize data extracted the from the selected papers[2]. To guide our analysis, we adopted three proposed taxonomies: one by Baldassarre et al. [4] to classify types of replication, one by Wohlin et al. [36] to analyze TTVs, and one by Easterbrook et al. [14] for types of empirical studies (see Section 2.1 for details).

[1] Some papers replicated more than one previous study.
[2] All data and scripts are available in our repository, see in Section ARTIFACT AVAILABILITY.

4 Results and Discussion

4.1 TTVs in Replication Studies: Overview

This section presents an overview of how TTVs are reported in ESE replication studies and their respective original studies. By analyzing the types of methods and the specific categories of TTVs addressed, we aim to understand to what extent replication efforts mirror or diverge from the validity concerns initially raised. The findings also provide insight into the types of validity threats prioritized in current research practices.

4.1.1 Study Selection and Dataset Overview. The SCOPUS search for the years 2022, 2023, and 2024 yielded 7,310 results[2] (see Figure 1). Based on the exclusion criteria (Section 3.2.1), 125 papers were selected after screening titles, abstracts, and keywords. From these, 41 papers were excluded because they did not include a clearly defined section, subsection, or highlighted area describing the types of TTVs. Additionally, one paper ([Rep10]) was removed because the original study it referenced could not be located. Although the authors cited their previous study as the original, it appeared that this study had not been published; the reference stated it was "to be submitted in June 2023". We contacted all the listed authors to request access to this work, but did not receive any responses. After these exclusions, 83 replication papers remained for analysis.

These replication papers referenced a total of 126 original studies. In some cases, a single replication paper referred to multiple original studies, for instance, [Rep52] replicated both [Ori52.1] and [Ori52.2], while [Rep32] included replications of [Ori32.1], [Ori32.2], [Ori32.3], and [Ori32.4]. In some cases, there are original studies that do not included a dedicated TTVs section. Nevertheless, since our primary focus is on the replication studies themselves, we did not exclude any papers based on whether the original studies reported TTVs.

4.1.2 Relationship Between Research Methods and TTVs. Table 1 presents the distribution of replication and original studies based on the types of methods used [14] and types of TTVs reported [36].

Table 1: Overview of Studies by Method and TTVs.

Types of method [14]	Replications (94*)	Originals (132*)
Controlled Experiment	56 (59.6%)	75 (56.8%)
Survey	15 (16.0%)	25 (18.9%)
Case Study	9 (9.6%)	13 (9.8%)
Action Research	1 (1.1%)	0 (0.0%)
Ethnography	0 (0.0%)	0 (0.0%)
Others	13 (13.8%)	19 (14.4%)
Type of TTVs [36]	**Replications (84*)**	**Originals (126)**
External	64 (76.2%)	72 (57.1%)
Internal	58 (69.0%)	63 (50.0%)
Construct	39 (46.4%)	51 (40.5%)
Conclusion	24 (28.6%)	25 (19.8%)

* The number of Original and Replication studies is greater than the number of papers because some papers report more than one research method and one paper includes two TTV sections (one for each method).

While one might expect replication studies to adopt the same research methods as their respective original studies, this is not always the case. In some cases, the replication used a different research method than the original study[3]. As presented in Table 1, the most common method used in both, the replication and original studies was *controlled experiments*, such as experiments and quasi-experiments, applied in 59.6% of the replication studies and 56.8% of the original studies. No studies employing *ethnography* were identified during the analyzed period. Overall, the results reveal a strong concentration of replications studies employing *controlled experiments* and a possible underrepresentation of some methodological approaches in TTV discussions.

A total of 13.8% of the replication studies and 14.4% of the original studies reported using methods not listed by Easterbrook et al. [14]. Studies in this category typically mention only general descriptors such as statistical, quantitative, or analytical analysis, without explicitly stating a methodological approach.

4.1.3 Types of Threats to Validity. Both replication and original studies reported the most threats related to *external* and *internal* validity. *External* validity threats were mentioned in 76.2% of replication studies and 57.1% of original studies, while *internal* validity threats appeared in 69.0% of replication studies and 50.0% of original ones. In contrast, threats to the validity of the *conclusion* were the least reported in both types of studies (28.6% in replications and 19.8% in original studies), which may reflect an underappreciation of this dimension or a lack of clarity on how to identify and report it. *Construct* validity threats, while less prominent than internal and external ones, were still notable, especially among original studies (51.0% in originals vs. 46.4% in replications).

Overall, the distribution of TTVs is quite similar between original and replication studies, reinforcing the idea that validity concerns are pervasive in empirical research regardless of the study's primary aim, whether to generate or verify evidence. The results suggest that researchers prioritize concerns related to experimental control (internal validity) and generalizability (external validity), regardless of the type of study. The relatively higher percentages observed in replication studies may reflect an increased sensitivity to these dimensions, possibly because replication efforts are more explicitly aimed at testing findings in different contexts and conditions.

4.1.4 Alignment Between Original and Replication Studies. We built 2×2 contingency tables for each TTV category by cross-tabulating its presence (1) or absence (0) in original studies versus their replication counterparts. For each table, we calculated the Matthews Correlation Coefficient (MCC) [10], a balanced measure of binary association, and applied a chi-squared test of independence (without Yates's correction) to assess the strength and significance of the association. We chose MCC because it is particularly suitable for imbalanced data, offering a more informative evaluation of association strength than accuracy or other simple metrics. The MCC values indicated moderate positive associations[4]: *Internal* (MCC = 0.352, $p = 0.0001$), *External* (MCC = 0.322, $p = 0.0003$), *Construct* (MCC = 0.444, $p < 0.0001$), and *Conclusion* (MCC = 0.402, $p < 0.0001$). In all cases, the chi-squared tests rejected the independence between the types of TTVs reported, demonstrating that each threat type

[3] For instance, the paper [Ori25] conducted a controlled experiment, while its replication [Rep25] used a case study.

[4] These results should be interpreted with caution due to the relatively small sample size of replication studies in our dataset.

reported in the original studies is significantly associated with its reporting in the respective replications. These results suggest that the reporting of specific types of TTVs in replication studies is moderately aligned with their reporting in the original studies, showing a tendency toward consistency in addressing these concerns.

Summary. Our findings reveal that replication studies not only frequently report threats to internal and external validity but also tend to reflect the same types of validity concerns highlighted in the original studies they replicate. Although some methodological diversity exists, particularly in the "Others" category, controlled experiments remain dominant. The statistical analyses confirm a moderate but significant alignment between the types of validity threats addressed in original and replication studies, suggesting a growing awareness and effort to maintain consistency in threat reporting across study generations.

RQ1: What is the current state of TTV reporting in both replication studies and the original studies they replicate? TTV reporting in both replication and original studies centers mainly on internal and external validity threats, while conclusion validity remains the least addressed. Despite some variation in methods, replication studies show moderate alignment with original studies in the types of TTVs reported, suggesting a shared concern with research validity across both study types.

4.2 Research Methods and Associated Types of TTVs in Replication Studies

Understanding how different research methods can influence the reporting of TTVs in replication studies is important to evaluate the methodological rigor and transparency of research. In this section, we analyze whether certain study designs are more likely to be associated with specific types of TTVs. By examining the distribution of TTVs across methods and over time, and their relationship with replication types, we aim to uncover patterns that may inform best practices for designing and reporting replications in ESE.

4.2.1 Statistical Associations Between Methods and TTVs. We examined associations between research methods and reported TTVs in both original and replication studies using chi-squared tests[2]. No statistically significant associations were found in original studies, although a borderline result was observed between *controlled experiments* and *construct* validity threats (p = 0.0860). In replication studies, significant associations[4] were found between the *survey* method and *conclusion* validity threats (p = 0.0230), as well as between *case studies* and the same threat category (p = 0.0313). This suggests that surveys and case studies tend to emphasize the validity of conclusions when replicated, while controlled experiments may more evenly cover all TTV categories. The absence of associations in original studies and their appearance in more recent replications may reflect growing attention to methodological transparency and validity reporting over time.

Thus, replication studies using *survey* or *case study* methods may place greater emphasis on threats to *conclusion* validity, in contrast to original studies, which did not show strong associations between method and TTV types. It is important to note, however, that replication studies were published more recently than the original studies, which may reflect an overall trend toward greater methodological rigor and increased awareness regarding the importance of reporting validity threats.

4.2.2 Temporal Trends in TTV Reporting by Method. The bubble scatter plot (Figure 3) shows that *controlled experiments* consistently report a high number of TTVs in all four categories, suggesting a stronger emphasis on methodological rigor. In contrast, *surveys* and *case studies* report fewer TTVs and often focus on specific types, such as *external* or *construct* threats. Less common methods, including *action research* and *ethnography*, appear infrequently in replications. These patterns have remained relatively stable over time for *controlled experiments*, while other methods display more variability and selective reporting. This visual evidence reinforces the notion that method choice strongly influences which TTVs are considered and reported.

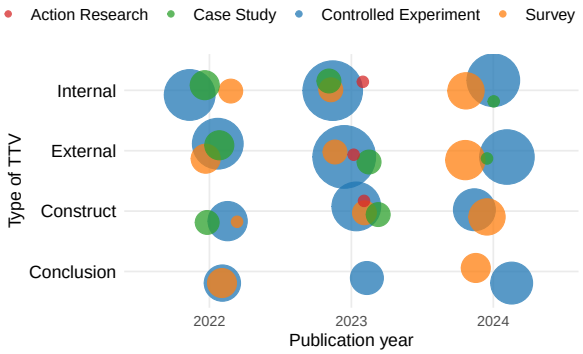

Figure 3: Mapping TTVs Reporting Patterns by Study Method Type and Year in Replication Studies.

4.2.3 Replication Types and Their Influence on TTV Reporting. Figure 4 shows that the most frequent replication types are *close* (30.1%), *differentiated* (24.1%), and *external* (22.9%), suggesting that researchers in ESE are engaging in both fidelity-focused replications, which aim to reproduce the original conditions, and variations that test the robustness of findings under different settings. *Conceptual* replications, which test the same hypotheses using different methods, appeared less frequently (16.9%), indicating a potential underuse of this strategy to explore original studies. *Internal* replications, repeated studies by the same research group, were the least reported (6.0%), suggesting limited self-replication efforts in the field. These results suggest that while validation and generalization are pursued, theoretical robustness and reproducibility are addressed with less intensity.

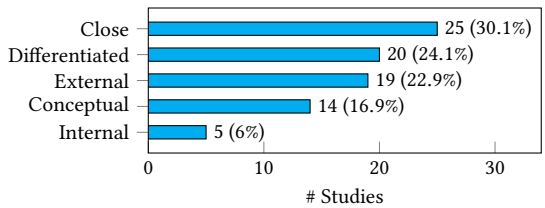

Figure 4: Frequency of replication types [4].

4.2.4 Replication Types and the Relationship to Methods and TTVs. The heatmap in Figure 5 adds nuance by linking replication types to both methods and TTV categories. It shows that *close*, *external*, and *differentiated* replications are the most common across methods and TTVs, reinforcing a trend toward replicating either with high fidelity or in new contexts. On the other hand, *internal* and *conceptual* replications are less frequent. The low number of *conceptual* and *internal* replications may indicate a missed opportunity to explore the theoretical generalizability and internal validity of previous findings. This pattern suggests that researchers conducting replications tend to prefer approaches that aim to reproduce results under conditions similar to the original study (close) or in different settings (external), rather than testing the robustness of conclusions (conceptual) or isolating specific internal elements (internal). These trends remain consistent across methods and TTV types, highlighting a replication landscape that prioritizes methodological fidelity at the expense of broader theoretical exploration and internal scrutiny.

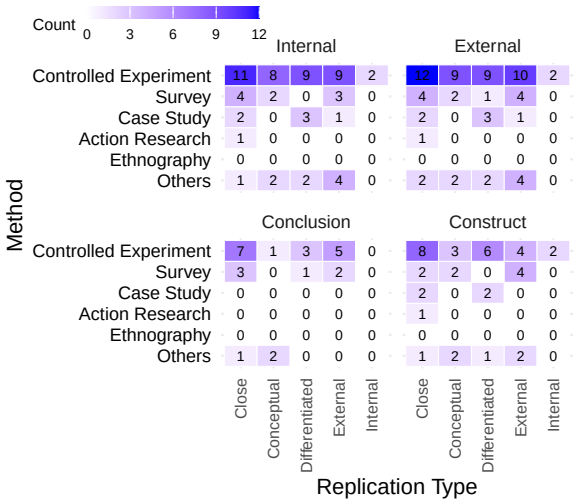

Figure 5: Replication Types by Method and TTV Category.

Summary and implications. Our analysis reveals that the reporting of TTVs in replication studies varies by research method and replication type. Controlled experiments consistently address most TTV categories, while methods such as surveys and case studies tend to emphasize specific TTV categories, such as the validity of conclusions. Moreover, the predominance of close and external replications suggests a focus on methodological fidelity and generalizability, whereas conceptual and internal replications remain underutilized. These findings highlight both progress and gaps in how replication studies report TTVs, pointing to opportunities for more diverse and reflective replication strategies in ESE research.

RQ2: What is the relationship between research methods and the types of TTVs reported in replication studies? There is a partial relationship between research methods and the types of TTVs reported in replication studies. Surveys and case studies show a significant association with conclusion validity threats, while controlled experiments tend to report all four TTV types more consistently. This suggests that certain methods emphasize specific validity concerns, whereas experiments support broader TTV reporting.

5 Concluding Remarks

This study investigated how TTVs are reported in replication studies within ESE, focusing on their frequency, specificity, and relationship to research methods and replication types. Compared to original studies, replication efforts demonstrate increased attention to TTVs, particularly for external and internal validity. However, construct and conclusion threats remain underreported. Controlled experiments consistently address all TTVs types, while surveys and case studies focus more narrowly. Most replications are close or differentiated, with conceptual and internal replications underused. Strengthening replication practices will require broader adoption of diverse replication types and more systematic reporting of TTVs. Together, these results highlight a growing but still uneven maturity in the design and reporting of replication studies in ESE. While there is clear evidence of improvement in how TTVs are handled, there remains a need for broader adoption of diverse replication types and for more comprehensive coverage of all validity dimensions. Future work should encourage the development of practical tools, such as standardized checklists and reporting protocols, that promote the systematic identification and mitigation of TTVs.

Planned Next Steps: Building on the results of this SLR, we intend to answer the research questions RQ3 and RQ4 through a qualitative and comparative analysis of replication studies, focusing on how TTVs are identified, reinterpreted, and addressed. The goal is to understand how the study setting influences the reporting of threats, assess the continuity of TTV handling between original and replication studies, and investigate whether replications contribute methodological value by revealing new threats not identified in the original studies. As a result of this, we will develop a preliminary guide for conducting replication studies from a TTV perspective.

Future Work: Future research should consider expanding the search string with additional relevant keywords to achieve broader and more comprehensive coverage of studies published within the Software Engineering domain.

- **RQ3** - How do replication studies address and compare the TTVs of the original studies?
- **RQ4** - Are the threats mentioned in the original studies also recognized in the replications, and do the replications identify new TTVs that were not present in the original studies?

Threats to Validity: Due to space constraints and the nature of this paper, at this stage, our TTVs are available only in our repository[2].

ARTIFACT AVAILABILITY

Artifacts available on Zenodo: DOI 10.5281/zenodo.15511661.

ACKNOWLEDGMENTS

This work is partially supported by INES.IA, CNPq grant 408817/2024-0. Sergio Soares is partially supported by CNPq grant 306000/2022-9. Ivanildo Azevedo is partially supported by the CAPES.

REFERENCES

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

[2] Apostolos Ampatzoglou, Stamatia Bibi, Paris Avgeriou, Marijn Verbeek, and Alexander Chatzigeorgiou. 2019. Identifying, Categorizing and Mitigating Threats to Validity in Software Engineering Secondary Studies. *Information and Software Technology* 106 (02 2019). doi:10.1016/j.infsof.2018.10.006

[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] 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

[5] Márcio Barros and Arilo Neto. 2011. Threats to Validity in Search-based Software Engineering Empirical Studies. *RelaTe-DIA* 5 (01 2011).

[6] Marvin Muñoz Barón, Marvin Wyrich, Daniel Graziotin, and Stefan Wagner. 2023. Evidence Profiles for Validity Threats in Program Comprehension Experiments. In *Proceedings of the 45th International Conference on Software Engineering* (Melbourne, Victoria, Australia) *(ICSE '23)*. IEEE Press, 1907–1919. doi:10.1109/ICSE48619.2023.00162

[7] 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

[8] Stefan Biffl, Marcos Kalinowski, Fajar Ekaputra, Amadeu Anderlin Neto, Tayana Conte, and Dietmar Winkler. 2014. Towards a semantic knowledge base on threats to validity and control actions in controlled experiments. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Torino, Italy) *(ESEM '14)*. Association for Computing Machinery, New York, NY, USA, Article 49, 4 pages. doi:10.1145/2652524.2652568

[9] Dante Carrizo and Jacqueline Manriquez. 2016. Assessment method of empirical studies in software engineering. In *2016 35th International Conference of the Chilean Computer Science Society (SCCC)*. 1–12. doi:10.1109/SCCC.2016.7836001

[10] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 1 (2020). doi:10.1186/s12864-019-6413-7

[11] 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/

[12] Margarita Cruz, Beatriz Bernárdez, Amador Durán, José A. Galindo, and Antonio Ruiz-Cortés. 2020. Replication of Studies in Empirical Software Engineering: A Systematic Mapping Study, From 2013 to 2018. *IEEE Access* 8 (2020), 26773–26791. doi:10.1109/ACCESS.2019.2952191

[13] Liliane da Silva Fonseca, Eudis Oliveira Teixeira, and Sergio Soares. 2019. ValiDE-Plan — Validity-Driven Software Engineering Experiments Planning Tool. In *Anais Estendidos do X Congresso Brasileiro de Software: Teoria e Prática* (Salvador). SBC, Porto Alegre, RS, Brasil, 102–107. doi:10.5753/cbsoft_estendido.2019.7665

[14] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. *Selecting Empirical Methods for Software Engineering Research*. Springer London, London, 285–311. doi:10.1007/978-1-84800-044-5_11

[15] Edison Espinosa, Juan Ferreira, and Henry Chanatasig. 2018. *Using Experimental Material Management Tools in Experimental Replication: A Systematic Mapping Study*. 252–263. doi:10.1007/978-3-319-73450-7_25

[16] Larissa Falcão and Sergio Soares. 2021. Human-Oriented Software Engineering Experiments: The Large Gap in Experiment Reports. In *Proceedings of the XXXV Brazilian Symposium on Software Engineering* (Joinville, Brazil) *(SBES '21)*. Association for Computing Machinery, New York, NY, USA, 330–334. doi:10.1145/3474624.3474649

[17] Andrea Fasciglione, Maurizio Leotta, and Alessandro Verri. 2022. Reproducibility in Activity Recognition Based on Wearable Devices: a Focus on Used Datasets. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 3178–3185. doi:10.1109/SMC53654.2022.9945344

[18] Robert Feldt and Ana Magazinius. 2010. Validity Threats in Empirical Software Engineering Research - An Initial Survey. 374–379.

[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] Lucas Gren. 2018. Standards of validity and the validity of standards in behavioral software engineering research: the perspective of psychological test theory. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Oulu, Finland) *(ESEM '18)*. Association for Computing Machinery, New York, NY, USA, Article 55, 4 pages. doi:10.1145/3239235.3267437

[21] 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

[22] Patricia Lago, Per Runeson, Qunying Song, and Roberto Verdecchia. 2024. Threats to Validity in Software Engineering – hypocritical paper section or essential analysis?. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Barcelona, Spain) *(ESEM '24)*. Association for Computing Machinery, New York, NY, USA, 314–324. doi:10.1145/3674805.3686691

[23] Ruchika Malhotra and Megha Khanna. 2018. Threats to validity in search-based predictive modelling for software engineering. *IET Software* 12, 4 (Aug. 2018), 293–305. doi:10.1049/iet-sen.2018.5143

[24] 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

[25] Kai Petersen and Cigdem Gencel. 2013. Worldviews, Research Methods, and their Relationship to Validity in Empirical Software Engineering Research. In *Proceedings of the 2013 Joint Conference of the 23nd International Workshop on Software Measurement (IWSM) and the 8th International Conference on Software Process and Product Measurement (IWSM-MENSURA '13)*. IEEE Computer Society, USA, 81–89. doi:10.1109/IWSM-Mensura.2013.22

[26] 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

[27] Adrian Santos, Sira Vegas, Markku Oivo, and Natalia Juristo. 2021. Comparing the results of replications in software engineering. *Empirical Softw. Engg.* 26, 2 (March 2021), 41 pages. doi:10.1007/s10664-020-09907-7

[28] Moritz Schloegel, Nils Bars, Nico Schiller, Lukas Bernhard, Tobias Scharnowski, Addison Crump, Arash Ale-Ebrahim, Nicolai Bissantz, Marius Muench, and Thorsten Holz. 2024. SoK: Prudent Evaluation Practices for Fuzzing . In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1974–1993. doi:10.1109/SP54263.2024.00137

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

[30] Janet Siegmund, Norbert Siegmund, and Sven Apel. 2015. Views on internal and external validity in empirical software engineering. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1* (Florence, Italy) *(ICSE '15)*. IEEE Press, 9–19.

[31] Fabio Q. Silva, Marcos Suassuna, A. César França, Alicia M. Grubb, Tatiana B. Gouveia, Cleviton V. Monteiro, and Igor Ebrahim Santos. 2014. Replication of empirical studies in software engineering research: a systematic mapping study. *Empirical Softw. Engg.* 19, 3 (June 2014), 501–557. doi:10.1007/s10664-012-9227-7

[32] Martín Solari, Sira Vegas, and Natalia Juristo. 2018. Content and structure of laboratory packages for software engineering experiments. *Information and Software Technology* 97 (2018), 64–79. doi:10.1016/j.infsof.2017.12.016

[33] Eudis Teixeira, Liliane Fonseca, Bruno Cartaxo, and Sergio Soares. 2019. PrioriTTVs: A process aimed at supporting researchers to prioritize threats to validity and their mitigation actions when planning controlled experiments in SE. *Inf. Softw. Technol.* 115, C (Nov. 2019), 20–22. doi:10.1016/j.infsof.2019.07.008

[34] 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

[35] Roberto Verdecchia, Emelie Engström, Patricia Lago, Per Runeson, and Qunying Song. 2023. Threats to validity in software engineering research: A critical reflection. *Inf. Softw. Technol.* 164, C (Dec. 2023), 4 pages. doi:10.1016/j.infsof.2023.107329

[36] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Böjrn Regnell, and Anders Wesslén. 2000. *Experimentation in Software Engineering: An Introduction*. Springer New York, NY. doi:10.1007/978-1-4615-4625-2

[37] Marvin Wyrich and Sven Apel. 2024. Evidence Tetris in the Pixelated World of Validity Threats. In *Proceedings of the 1st IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering* (Lisbon, Portugal) *(WSESE '24)*. Association for Computing Machinery, New York, NY, USA, 13–16. doi:10.1145/3643664.3648203