Views on Internal and External Validity in Empirical Software Engineering: 10 Years Later and Beyond

Alina Mailach
ScaDS.AI Dresden/Leipzig, Leipzig University
Leipzig, Germany

Janet Siegmund
Chemnitz University of Technology
Chemnitz, Germany

Sven Apel
Saarland University
Saarbrücken, Germany

Norbert Siegmund
ScaDS.AI Dresden/Leipzig, Leipzig University
Leipzig, Germany

Abstract

Ten years ago, a survey (ICSE'15) revealed an unsatisfactory picture among key players of the software engineering research community: There was dissent and confusion on how to approach empirical research. Specifically, researchers were divided on the tradeoff between internal and external validity, holding strong and often opposing opinions, which cannot be a basis for stringent scientific progress. Clearly, the field has progressed over the last decade, but did the views progress, as well, or is the community still divided and confused? Our study addresses these questions by replicating the original survey among current key players of the field. Analyzing 790 open answers learning how perspectives have changed over the past ten years, we found that, despite increased awareness of the intricacies of conducting empirical studies, not nearly enough has changed to address the wide range of opinions on what a good study is and how that can be reflected in the review process. Specifically, participants disagree on balancing internal, external, and ecological validity, and while there is consensus on the need for replication studies, the specifics of when, how, and what to replicate remain unclear. Our results suggest the need for a more sophisticated review process, incorporating clear empirical standards for various methods and fostering honest discussions on what is worth replicating.

CCS Concepts

• **Software and its engineering**;

Keywords

Validity, Software Engineering Research, Replication

ACM Reference Format:
Alina Mailach, Janet Siegmund, Sven Apel, and Norbert Siegmund. 2026. Views on Internal and External Validity in Empirical Software Engineering: 10 Years Later and Beyond. In *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE '26), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3744916.3764520

1 Introduction

A key question all researchers have to answer is how to gain valid insights and build trust in scientific results. Specifically, in empirical research, many decisions influencing the validity of the results are left to the researchers, introducing a degree of arbitrariness and a piece of doubt.

A central decision that entails most research endeavors is whether we want to aim for generalizable findings or validated causal statements. Naturally, we want to achieve both, but a plethora of confounding factors stemming from practical and complex settings, human preferences and experiences, as well as diverse environments and use cases, to name a few, often blur the picture. As a result, it is unclear whether we observe a result due to our treatment (i.e., the independent variables of an experiment or study that we vary) or by chance or any (unknown) combination of factors of the study.

The awareness and preference on this tradeoff and how it should be addressed was analyzed and published by Siegmund et al. [37] in ICSE 2015, in which three of the four authors were involved. A survey conducted among 79 program committee and editorial board members of major software engineering venues revealed that preferences vary greatly: Some researchers think the community should prioritize internal validity in research designs over external validity, whereas others advocate for the exact opposite. Yet other respondents in the survey showed no awareness of the tradeoff at all and suggest both to be maximized. The survey also found that, while most researchers are aware of the necessity of replication to establish trust in research results, there was no shared understanding on how replications should be approached. This situation was clearly unsatisfactory. Siegmund et al. concluded that having reviewers who do not fully understand the tradeoff between internal and external validity, yet have strong opinions on maximizing one over the other, turns the reviewing process into a game of chance, as papers are largely judged on personal preference instead of objective sound research design that values either direction.

Ten years have passed since this assessment of the community's views on the future of empirical software engineering. Yet, discussions and research on the validity of empirical research results are still ongoing [4, 27, 28, 38, 41, 44]. Even reviewer guidelines of major software engineering venues still lack an awareness of this issue despite the clear demonstration of strong conflicting expectations on research designs. Thus, it is time to reassess the community's views and identify the direction in which we are heading by asking these questions: Is the community now more aware of the tradeoff

ICSE '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2025-3/26/04
https://doi.org/10.1145/3744916.3764520

between internal and external validity? Do reviewers agree more on how to address this tradeoff? How has the community's perspective on the necessity and practice of replicating research results evolved, especially after seeing artifact reviews, replication and reproduction badges, and dedicated venues?

To answer these questions, we have replicated the survey conducted by Siegmund et al. and compare the views of 68 of today's key players in the software engineering community to those surveyed ten years ago. By examining these views, we are able to provide insights into the current trends and future directions in empirical software engineering research. Ultimately, we aim at reigniting and stimulating the discussion on how research studies should be designed and published and help the community in revising and defining standards.

Interestingly, during the analysis of the survey conducted ten years ago, the authors observed a conflation of the concepts external and ecological validity. External validity refers to the generalizability of research results to other settings and populations, whereas ecological validity concerns the applicability of results from the study setting to real-world settings, characterized by practicality. Although closely related, results from a study with high ecological validity (e.g., conducted with real-world developer teams from one company) are not necessarily externally valid (e.g., transferable to teams of other companies), and vice versa. Clearly, times are changing, and so are the discussions within the community. We want to assess whether the conflation of external and ecological validity still exists and, therefore, explicitly asked key players about their awareness and opinions on the distinction between these two concepts. This way, we aim at understanding the view of the community in greater detail and promote a more nuanced discussion of validity in empirical software engineering.

In summary, we make the following contributions:

- A survey conducted among 68 members of editorial boards and program committees of major software engineering research venues.
- A discussion of trends in perspectives among these members, comparing data from today and ten years ago.
- An assessment of current views and of the awareness and opinions on the distinction between external and ecological validity, which were found to be conflated in the survey ten years ago.
- Comprehensive supplementary material, including all data and resources to facilitate independent replication and reproduction of our results [22].

2 Background and Related Work

To ensure a common ground for this paper, we shortly introduce the concepts of internal, external, and ecological validity. For clarity, we refer to Siegmund et al.'s 2015 study as the *original* study and the present study as the *replication* study in the remainder of this paper.

2.1 Validity

For illustrating the different validity concepts, we use the same running example as in the original study: Suppose researchers want to determine whether a common construct, *func*, from functional programming (e.g., algebraic data types, pattern matching) is more comprehensible to programmers than a construct, *oops*, from object-oriented programming (e.g., inheritance, dynamic dispatch). They plan to conduct a study with human participants.

Numerous factors influence comprehensibility, including programming language, IDE support, familiarity of participants with languages, the problem to solve, and programming experience. While accounting for all these factors is crucial in empirical research, it is practically impossible in a single study. Doing so would require millions of participants [36]. Therefore, researchers must make several design decisions.

Researchers might choose to control for all these influences by designing an artificial programming language differing only in the constructs *func* and *oops*. In other words, researchers focus on *internal validity* to understand the influence between the choice of *func* or *oops* on comprehensibility. While being very rigorous and controlled, this limits our ability to understand how these constructs affect real-world programming.

Alternatively, researchers might focus on the effects of these constructs in real-world scenarios by having developers use Haskell for functional programming and Java for object-oriented programming. In other words, researchers prioritize *ecological validity*, obtaining observations in a practical setting. The crux is that both programming languages differ in more than the constructs *func* and *oops*: They also differ in syntax, programming concepts of modularization, and canonical coding guidelines. Thus, any observation about comprehensibility is potentially influenced by all these differences, as well, and researchers would not be able to distill to what extent comprehensibility is influenced by the constructs *func* and *oops* or other differences of the programming languages.

A third choice is to aim at *external validity*, which describes to what extent the observation in a study can be applied to different contexts. For example, if the participants work with Haskell and Java as programming languages, to what extent would the observations also be applicable for Scheme, Lisp, C++, or C#? A study with high external validity could include all these programming languages, so that observations are also obtained for these other programming languages.

So researchers face a tradeoff: including more factors makes it harder to identify causal effects of *func* and *oops*, while controlling for more factors makes it harder to generalize results. Additionally, the practicality of the study setting (ecological validity) affects the findings. While a study might increase external validity without affecting ecological validity (as in our example by adding more programming languages), this is not always the case. For example, a study in a single company with a specific language and system may be highly ecologically valid (and possibly internally valid), but results may not generalize to any other company or programming language. Thus, researchers must often make design decisions that balance different kinds of validity. Ten years ago, as empirical software engineering was transitioning from a niche to an established field and community, Siegmund et al. set out to determine whether the knowledge of the community kept pace with the growing numbers of empirical studies.

2.2 Related Work

Next, we concentrate on related work published after the study of Siegmund et al. For an overview of related work published prior to 2015, we refer to the related work section of the original study [37]. Closest to our study is a study conducted by Galster and Weyns [15], in which they repeat parts of Siegmund et al.'s questionnaire in the software architecture community. The results indicate that, even in a more narrow subfield, there is no agreement on how the tradeoff between internal and external validity should be approached, no consensus on the usefulness of student samples, and replications are seen as valuable but rarely published, reflecting some insights of the original study. While our replication study additionally explores the conflation of external and ecological validity, Galster and Weyns focus more on general attitudes toward empirical research and preferences for quantitative or qualitative research, observing no agreement within the community on these aspects, either.

In addition, there is a growing number of valuable guidelines on methodologies in software engineering [39, 40]. Unlike these, our study focuses on perspectives and opinions, providing a base for future development and refinement of such guidelines.

Validity in software engineering research. Most research targeting validity in software engineering focuses on addressing threats to validity and reporting them from various perspectives. Wyrich and Apel [44] advocate for evidence-based methods over relying on researchers' intuition. Other work suggests framing study limitations as trade-offs, providing rationales for decisions, and discussing alternatives not chosen [28]. Verdecchia et al. criticize the "Laundry List" approach to validity threats, arguing that merely listing threats can hinder meaningful discussion of design trade-offs [41]. Sjøberg and Bergersen [38] specifically focus on improving the reporting of construct validity threats.

Several approaches have been developed to collect, analyze, and consolidate threats to validity in various subfields of software engineering, such as secondary studies [1, 46], code comprehension [13], and software traceability models [23]. These studies provide researchers with an overview of potential threats and how to assess and mitigate them. Beyond reporting and collecting threats to validity, Ralph and Tempero focus on construct validity, providing a theoretical background and practical guidelines on how to assess it [27]. While all this work focuses on different aspects of reporting threats to validity and guiding researchers on how to assess and mitigate them, our study concentrate on researchers' views and preferences that guide design decisions in the first place. Our study further examines views on ecological validity, which has received limited attention in software engineering studies. This is in contrast to other fields, such as psychology [16, 25, 33], justice [7, 42], and education research [20], which debate ecological validity and its impact on results of empirical studies.

Replication in software engineering research. Replications in software engineering have gained significant attention from various researchers. Cruz et al. [10] found in a systematic mapping study that replications are becoming more common, with a significant focus on controlled experiments. However, not all subfields are equally publishing replications, and the quality of reporting in both, original and replication studies, is often insufficient [35].

Others identify challenges in replications [12] and provide guidelines and support on how to compare results of replications [30–32]. Meta-analytical methods are found to be inferior compared to other analysis strategies [30]. In this vein, Shepperd [34] conducted a simulation study and finds that replicating under-powered studies results in wide prediction intervals, leading to most results of a replication being evaluated as confirming the original results. These studies provide valuable methodological considerations. In contrast, our study does not seek to improve the practice of replication, but aims at understanding their role and relevance within our community.

3 Methodology

3.1 Research Questions

We repeat the original study conducted ten years ago. Our goal is to understand current views of the community regarding empirical research and to identify trends and changes over the past decade. Thus, we ask the same questions again:

RQ$_1$ How aware is the research community of the tradeoff between internal and external validity?

RQ$_2$ What does the research community think on how the tradeoff between internal and external validity should be addressed?

RQ$_3$ How does the research community see the role of replication?

These questions directly address key aspects of empirical methodology [43], enabling us to identify trends by comparing answers of current reviewers with those from ten years ago.

As our understanding and the issues that are open have evolved over the past decade, we added a new research question on the perception of the community on ecological validity:

RQ$_4$ How does the research community think about ecological validity?

By answering these questions, we provide a more nuanced discussion, shedding light on unexplored aspects of validity. Clearly, empirical research extends beyond these concepts, and other aspects might be equally worth exploring. However, to understand how the community has changed over the last decade, we focus on the concepts highlighted in the original study (internal/external validity and replication). We additionally include ecological validity, as this has been often conflated with external validity.

3.2 Survey Questionnaire

We used a revised and extended version of the original questionnaire (including definitions and a scenario) and collected data using LimeSurvey, hosted by Leipzig University. At the beginning of the questionnaire, we introduced the participants to definitions of internal and external validity and asked for information on their reviewing activity in the past four years. We then presented an example research scenario and asked specific questions. All questions and answer options are listed in Table 1. Questions marked with ⋆ are new and were not part of the original study.

Scenario. We provided participants with the same scenario of the original questionnaire (cf. Section 2.1, *fun* vs. *oops*) and presented two options how a study of a paper submitted for review might approach this research goal, with the first option maximizing

Table 1: Mapping of survey questions to corresponding research questions.

RQ	Question	Answer options
1,2	Which option would you prefer for an evaluation? [This question was asked twice, once for the human scenario and the non-human scenario]	□ Maximize int. validity □ Maximize ext. validity □ No preference
1	Would it be a reason to reject a paper that does not choose your favorite option?	□ Yes □ No
1,2	In your opinion, what is the ideal way to address research questions like the one outlined above?	Open
1	Did you recommend to reject a paper in the past mainly for the following reasons?	□ Int. validity too low □ Ext. validity too low
1,2	For research questions like the one presented above (common functional vs. object-oriented language constructs), do you prefer more practically relevant research or more theoretical (basic) research?	□ Applied research □ Basic research
1	During your reviewer career, have you changed how you judged a paper regarding internal and external validity?	□ Yes □ No
1,2	Do you have any suggestions on how empirical researchers can solve the dilemma of internal vs. external validity?	Open
2	[Asked once with and without humans] In your opinion, do you think that in the literature, empirical evaluations with/without human participants...	
	...are needed more or less often?	□ Considerably more often □ More often □ Fine as is □ Less often □ Considerably less often
	...are accepted/rejected too often?	□ Considerably too often rejected □ Too often rejected □ Fine as is □ Too often accepted □ Considerably too often accepted
	...need higher internal/external validity?	□ Considerably higher int. □ Higher int. □ Fine as is □ Higher ext. □ Considerably higher ext.
3	Do you think we need to publish more experimental replications in computer science	□ Yes □ No
3	During your activity as a reviewer, how often have you reviewed a replicated study?	□ Never □ Rarely □ Sometimes □ Regularly
3	In general, how were the replications rated by you/by your fellow reviewers	□ Accept □ Borderline □ Reject □ Not applicable
3	As a reviewer of a top-ranked conference, would you accept a paper that, as the main contribution,...(assuming authors realized it in the best possible way)	
	... exactly replicates an experiment of the *same/another* research group?	□ Yes □ No □ I do not know
	... replicates an experiment of the *same/another* research group but *increases int. val.*?	□ Yes □ No □ I do not know
	... replicates an experiment of the *same/another* research group but *increases ext. val.*?	□ Yes □ No □ I do not know
3	During your activity as a reviewer, did you notice a change in the number of replicated studies?	□ Yes, it increased □ Yes, it decreased □ No
4	★ What do you think the community can do to recognize and reward replications, such that conducting and publishing them becomes more attractive?	Open
5	★ Did you notice situations in which the concepts external and ecological validity were mixed up?	□ Yes □ No □ I do not know
5,6	★ Do you think the software engineering research community should be more aware of the distinction?	□ Yes □ No □ I do not know
6	★ Do you think that the software engineering research community should strive for more ecological validity?	□ Yes □ No □ I do not know
6	★ ... [if yes] Which of the following would be acceptable to compromise on, when conducting more ecologically valid studies?	□ internal □ external □ replicability

internal validity and the second option maximizing external validity. This was followed by another example of a study that did not recruit human participants, but aims at evaluating a new method that promises faster response times for Web applications. Again, we presented two options.

Closed and open questions. Our questionnaire contained 22 closed questions, allowing participants to choose one or more predefined options. For each closed question, participants were optionally asked to elaborate using an open text field. At the end of the questionnaire, participants got the opportunity to share final thoughts on each covered topic as well as the overall questionnaire.

Modifications to the original questionnaire. We made minor adjustments to the phrasing and placement of the scenario and modernized terms (e.g., replacing *SourceForge* with *GitHub* or *databases* with *web applications*). Additionally, we changed the wording in the presented research study maximizing external validity, such that it does not explicitly target the creation of a practical, everyday setting, but rather a setting that captures the essence of many, possibly realistic settings. The original wording might had influenced participants to associate external validity with ecological validity, and we adjusted this to mitigate this threat. Besides that, we did not change general wordings, including some ambiguities, since the primary goal of our study is to identify trends and compare answers from key players today with those from ten years ago.

New questions. To answer our novel research question RQ$_4$, we added four questions at the end of the questionnaire to ensure the first part of the questionnaire remains as close to the original survey as possible. We first provided participants with a definition of the term ecological validity and explained that the concept is related to external validity, but distinct. This ensured that participants could answer the corresponding questions, even if they had never encountered the term before. Additionally, participants were not allowed to go back and change their answers based on the concrete definition of ecological validity. This allowed us to compare answers to our questionnaire to the original study.

3.3 Study Setup

Participants. To gather insights from current leaders in software engineering, we collected the names and email addresses of program committee members from the following nine conference websites for the years 2021-2024:

- ASE (Software Engineering)
- EASE (Empirical Software Engineering)
- ECOOP (Programming Languages)
- ESEC/FSE (Software Engineering)
- ICPC (Program Comprehension)
- ICSE (Software Engineering)
- ICSME (Software Engineering)
- OOPSLA (Programming Languages), and
- ESEM (Empirical Software Engineering).

Additionally, we included the current editorial board members of the major journals TSE, EMSE, and TOSEM as of Dec. 2023.

A total of 68 participants provided an answer regarding their validity preference, which is similar to response rates to other surveys in our field [6, 21, 45]. On average, participants were part of 9.3 (±6.8) program committees or editorial boards in the past four years.

Quantitative analysis. For each unchanged closed question, we tested whether the distribution of responses of our participants significantly differs from that in 2015. That is, we performed exactly one statistical test per question, with the null hypothesis being: *There is no significant difference in the distributions between 2015 and 2025.* We used Pearson's χ^2 -Test for independence if the number of participants satisfies the prerequisites of the test (i.e., expected frequency above 5). Otherwise, we used Fisher's Exact test, which is more robust for smaller sample sizes [2].

Qualitative analysis. We employed open card sorting [17, 47] to identify shared themes across the answers of participants to open questions. Although the categories from 2015 are available, we chose open card sorting to avoid biasing our analysis and to allow for the discovery of new themes. At the same time, we were familiar with the original categories, so we did not approach the data entirely uninformed. To balance both concerns, we conservatively began with open coding and later compared our categories to those from the original study.

The open card sorting was conducted by three researchers over multiple sessions (2–3 hours, each). For each survey question, we wrote responses on individual cards, one response per card. Starting with the first card, each researcher read the response, and the team discussed and assigned it to a category. Subsequent cards were either assigned to existing categories or led to the creation of new ones. If a response addressed multiple themes, we physically split the card and categorized the parts separately. After sorting all responses for a question, we reviewed the categories and assignments to ensure consistency. Finally, we compared our categories to those of the original study. For any insight not found in the categories 10 years ago, we revisited the raw data from 2015 to check for similar motives, to ensure the themes were, in fact, not present back then.

4 Results

For each question, we provide an overview of responses to closed questions, alongside the results of the original study, followed by results from our qualitative analysis. We then answer each research question. In total, we spent 37 hours analyzing 790 open answers. We provide a deeper discussion and comparison of results to the original study in Section 5.

4.1 RQ$_1$: How aware is the research community of the tradeoff between internal and external validity?

4.1.1 Qualitative results. For this research question, we directly provide qualitative results, as there were no closed questions to answer. We make two observations: (1) the community is aware of the tradeoff between internal and external validity, and (2) there is a large spectrum of methodological preferences beyond quantitative empirical research.

Awareness of the tradeoff. Participants in our replication study generally exhibit awareness of the tradeoff between internal and external validity, often stating that these should be balanced or that maximizing either is equally valuable:

> *"[..] find a good compromize, not go to the extreme in one direction and ignore the other."*$_{P35}$.

We received no statements indicating that participants deny or neglect this tradeoff. Overall, this indicates a recognition and awareness of the complementary nature of internal and external validity by our study participants.

Beyond quantitative empirical research. Several participants mention that they prefer qualitative or mixed-method studies over quantitative studies and controlled experimental designs, for example:

> *"I believe it [the research question in the provided scenario] probably can't be addressed by a controlled trial on human participants. Instead the best you could do is probably longitudinal study using ethnographic methods [..]"*$_{P23}$.

We find this sentiment across several questions and participants; for example, one participant indicates that internal and external validity might not be suitable concepts to judge qualitative research:

> *"I don't really use the terms [int. and ext. validity] myself, but prefer quality criteria used in qualitative research"*$_{P43}$.

Finally, one participant mentions that the rejection of papers based on methodological preferences is a key problem:

> *"A far more severe problem is the lack of understanding in the community of reviewers who reject, for example, qualitative research papers based on a matter of taste [..]"*$_{P53}$.

4.1.2 Discussion of RQ$_1$. The results of our replication study indicate that, while there is awareness of the tradeoff between internal and external validity, the distinction and confusion in quality assessment of different empirical methods, including quantitative and qualitative studies, is an increasing concern. The original survey had two concrete research scenarios in which the role of different kinds of validity can be discussed. However, the survey assumed more or less a quantitative setting for this discussion. Multiple answers raised the issue that such research scenarios can also be answered with qualitative methods, one participant stated that they

> *"find this [focus on quantitative empirical research under the term 'empirical research'] a serious bias against qualitative empirical research, which I find is a serious problem, as it hinders empirical software engineering researchers to address, among others, the trade off discussed here."*$_{P7}$.

We fully agree: (i) addressing and discussing validity should not follow a blind automatism, as in cases in which

> *"authors tend to copy things [in the validity section] from one paper to the other without putting much thought into it."*$_{P53}$.

And (ii) respect the empirical method followed as participants indicated, that qualitative studies might get rejected due to taste. So, while we focus on validity aspects that are more pronounced in quantitative empirical research, similar problems regarding lack of awareness of validity aspects, tradeoffs, non-applied concrete reviewing guidelines, and diverging reviewer expectations exist for qualitative research and need more attention.

Key Results RQ$_1$

- Participants understand and accept the tradeoff between internal and external validity, indicating community awareness.
- Several participants indicate that they are generally more concerned with a lack of understanding of different empirical research methods.

4.2 RQ$_2$: What does the research community think on how the tradeoff between internal and external validity should be addressed?

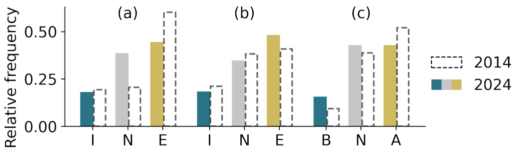

Figure 1: Closed questions for RQ$_2$: Which option would you prefer? I=Maximize internal; E=Maximize external; N=No preference; (a) Scenario with human participants (b) Scenario without human participants (c) Which do you prefer? B=Basic research (theoretical, knowledge-driven); N=No preference; A=Applied research (practical, solution-oriented)

4.2.1 Quantitative results. Figure 1, shows responses of participants to the closed questions in our replication survey regarding RQ$_2$, and for comparison also the results from the original survey. Similar to the prevalent opinion in 2014, participants favor external validity over internal validity. However, this preference has slightly diminished in favor of a more balanced view. Overall, the preferences observed in 2014 remain stable, with none of the differences in distributions being significant at $\alpha = 0.05$, according to a χ^2-test; detailed p-values and test statistics for all questions are provided in the supplementary material.

4.2.2 Qualitative results. Many participants state that there is no ideal way to perform empirical studies. Nevertheless, they provide different perspectives on how to approach the tradeoff: Participants recommended to compromise between internal and external validity, to maximize one at the cost of the other, or to conduct a family of studies, first focusing on internal validity, then on external. Some participants recommended to dismiss the research questions provided in the scenario (whether functional or object-oriented constructs are more comprehensible) as unworthy of investigation.

Reasons to prioritize external validity. The most prevalent reason to favor external validity is that insights from an externally valid study are considered more useful or practical and relevant for industry:

> *"[..] the second one appears to be more likely to have 'practical' findings."*$_{P12}$,

> *"I work closely with industry and therefore value external validity more"*$_{P42}$.

Several participants stated that a too artificial setting (often needed for controlling confounding factors) has little relevance in practice, for example:

> *"[..] [in the internally valid study] the environment is that 'artificial' that it may be far away from reality."*$_{P50}$,
>
> *"I would prefer a realistic evaluation."*$_{P57}$.

Such statements appear in several responses, indicating that, generally, external validity seems to be associated with a realistic setup and practical relevance of the findings. However, this is a misconception: A realistic setting may not be generalizable, as it can be highly specific to a single company and may not apply to other company settings. In such cases, a realistic setting has low external validity.

Reasons to prioritize internal validity. Notably, participants mention that there are too many uncontrollable factors when maximizing external validity, such that there is

> *"[..] high risk of just regurgitating existing biases within the community [..]"*$_{P18}$.

In the same vein, some participants who prefer prioritizing internal validity state that internal validity is necessary to ensure that the study is correctly focused:

> *"It [maximizing internal validity] would ensure the focus is on the constructs we are interested in."*$_{P37}$,

> *"[..] without precise control of internal validity, it is not possible to argue why the study design addresses it [human understanding] in the first place [..]"*$_{P20}$.

The sentiment here is that, only with control, researchers can answer questions in a valid way.

Multiple studies with varying tradeoffs. Several participants state that researchers should perform multiple studies on the same topic and address similar or the same questions. Most prevalent, participants recommend to first perform studies with a focus on internal validity and then aim for generalizability with more externally valid studies:

> *"Starting with a highly-controlled study and then extending generalizability gradually with wider and wider studies would lead to a strong theory here."*$_{P64}$.

In this sense, answering a research question may not be a one-time experiment (or paper), but more in terms of a larger research program.

Depends on goal and involvement of humans. Several participants mention that whether to aim at internal or external validity depends on several aspects, such as the goal or rationale of the study. Further, it also depends on whether the study involves humans or is purely technical. For instance, one participant favors internal validity for human studies and external validity for technical studies, since

> *"This [non-human study] is an easier experimental design since it involves direct measurement [..]"*$_{P30}$.

This answer indicates that different criteria to human and non-human studies apply. This is in contrast to participants stating that, for both, the same criteria apply:

> *"I think from a scientific point of view, the problem is basically the same."*$_{P44}$.

Interestingly, even the participants who would apply different criteria for human and non-human studies do not agree on whether the human scenario involves more or less factors that need to be controlled for.

Reasons for rejection. When asked whether participants would reject a study based on their personal validity preferences, participants indicate that the choices a study makes on validity are not by themselves a reason for rejection, but that rejection rather depends on the quality of the study, including a reasonable discussion of threats to validity and rationales provided for design decisions. Only few indicate that they are likely to reject a paper if it does not choose to maximize internal validity, because

> *"Favoring external validity for a novel research question may miss out on important confounding [factors] and effects"*$_{P22}$.

Controversy over human participants. We find strong and partly opposing views on which population should be recruited from to conduct a study. Most prevalent, several participants would oppose recruiting students as participants:

> *"I'd rather not bid on a paper that uses students to evaluate features of programming languages."*$_{52}$.

By contrast, one participant would recommend the rejection of a study that does not follow their favorite option, and elaborates that they are

> *"very sceptical of using volunteers without control and especially open-source projects to answer general questions [..]"*$_{P20}$.

4.2.3 Discussion of RQ$_2$. There are diverse views on how the trade-off between internal and external validity should be approached, with many participants favoring external validity. Again, a common misconception is that external validity represents a realistic setting, resembling the everyday life of developers, while internal validity is associated with artificiality, such as in highly controlled lab studies. Similarly, the usefulness and impact of a study for industry is often associated with external validity:

> *"I favor relevance for developers over pure academic questions"*$_{P39}$.

However, some participants believe that industry actors, rather than academics, should determine a study's usefulness:

> *"In the past I worried industry and practicality had no voice, but increasingly I feel the responsibility for assessing practical impacts should be in the hands of industry."*$_{P25}$.

This way, the participant challenges the assumption that realistic studies are inherently more impactful.

The variety in perspectives among participants reveals several fundamental disagreements: from rejecting papers based solely on methodological considerations to concrete results as a determining factor; from broad research questions to narrower, more focused questions to not investigating certain questions at all; and from conducting studies with students as participants to only professionals being suitable participants. While differing opinions are not a problem per se, the crux is that, during the peer review process, a single paper is likely judged based on these opposing opinions. Current reviewing guidelines, such as those provided by conferences [11], do not typically address these diverse preferences, leaving reviewers and PC chairs (and authors) unsupported. Establishing comprehensive guidelines or promoting existing ones that consider these differing perspectives could enhance the consistency and fairness of the peer review process. This may involve a fairer assignment of reviewers based on similar preferences, which we will discuss in Section 5.2.

Key Results RQ$_2$

- Generally, participants prefer external validity over internal validity.
- There is a conflation between the notion of external validity, the realism of the setting, and the practical impact of research results.
- Participants express a wide range of preferences regarding the assessment of a paper, often leading to fundamental disagreements on several aspects of conducting research in software engineering.

4.3 RQ$_3$: How does the research community see the role of replication?

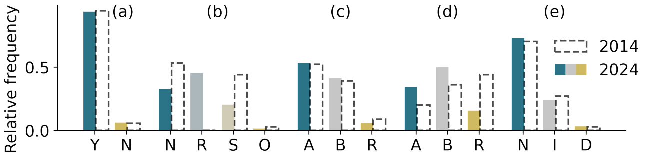

Figure 2: Closed questions for RQ$_3$: (a) Do you think we need to publish more experimental replications in computer science? Y=Yes; N=No; (b) How often have you reviewed a replicated study? N=Never; R=Rarely; S=Sometimes; O=Regularly; In general, how were the replications rated A=Accept; B=Borderline; R=Reject (2024 only); (c) ... by you; (d) ... by your fellows; (e) Did you notice a change in the number of replicated studies? N=No; I=Increased; D=Decreased

4.3.1 Quantitative results. In Figure 2, we show the answers of participants to the closed questions for RQ$_3$. Again, the distributions of answers of now and ten years ago are similar, and none of the differences is statistically significant: Most participants still believe more replications are needed, yet few regularly review them and still, participants feel like they are more likely to accept replications than their colleagues.

4.3.2 Qualitative results. Overall, most participants state that replications are valuable and mention several reasons why we need more of them. Among others, replication should be conducted to consolidate findings, increase external validity of studies, and ensure the progress of the field. Some participants mention that we should ignore results until replicated or even retract original studies that fail to replicate later. Participants repeatedly mention that, in the review process, the novelty criterion[1] is used to argue against acceptance of replications. Some indicate that

> *"[..] You currently have to be lucky to get good editors/reviewers who fight back against this outdated view."*$_{P6}$.

Others see the responsibility in the steering committees and editorial boards to change reviewing criteria.

How to replicate and when to accept replications. Participants express differing opinions on how replications should be conducted. Some value all replications, others have strong reservations against exact replications, where the study design is kept

[1] *i.e., "the extent to which the paper is sufficiently original with respect to state-of-the-art"* e.g., ICSE2023-25

the same, conducted by the same authors as the original study; one participant asked

> *"[..] isn't it just the same damn paper?"*$_{P11}$.

Additionally, participants raise concerns regarding replications conducted by the same authors, because they get suspicious,

> *"[..] especially if the results were similar."*$_{P46}$.

Or because same-author replications might

> *"[..] suffer from a file-drawer effect: we won't see the failures"*$_{P25}$.

Both views seem to express distrust in same-author replications. Several participants mention that they are willing to appreciate and accept replications when there is a certain delta between the original study and the replication that increases internal, external, or both kinds of validity.

Notably, several participants state that not all outcomes of replications are equally valuable, but that acceptance should depend on the *"[..] usefulness of the outcomes"*$_{P8}$, which can hardly be seen upfront. Some argue that confirming findings is less useful than showing contradictions, which contradicts the main purpose of replications on increasing trust of proposed methods and theories. Finally, several participants mention that not all work is equally worth replicating. For instance, one participant describes:

> *"[..] Most of the papers I read do not have high level findings [..] advocating for replicating experiments is a bit like putting the cart before the horse."*$_{P27}$.

Some participants suggest that subcommunities should collaboratively decide which research to replicate and then advocate for it, for instance, by organizing special events. Additionally, several participants indicate that we need clearer guidelines and support structures for all involved groups: for researchers conducting replications, for groups whose studies failed to be replicated, and for reviewers of replications.

Where to publish replications. There are different and contradictory opinions on how and where to publish replications: While some participants state that dedicated tracks and venues should be the platform for replications, others see replications as equally valuable as novel contributions, so they should be published in the main track of conferences avoiding to *"[..] ghettoize [..]"*$_{P25}$ them. Still others think that conferences are meant for novel work only.

Incentives for replications. Only two participants indicate that current efforts are sufficient or that *"this [promoting replications] is not a problem"*$_{P7}$. Several participants indicate that we should generate incentives for conducting replications via citations, badges, and awards. However, others think that

> *"[..] Badging is pretty useless in my view. [..]"*$_{P25}$.

So, participants do not agree on which incentives work and which do not. Another proposed idea is that replications are good teaching methods and could be included into the curriculum of graduate and PhD studies:

> *"[..] expect every PhD student to include at least one replication [..] in their thesis [..]"*$_{P45}$.

This way, the importance of replications could be emphasized already in early stages of the academic career. Finally, one participant mentions that a key problem is missing funding and acknowledgment of replications by funding agencies:

> *"[..] Who gives money for a research topic that somebody else already covered? [..]"*$_{P43}$.

4.3.3 *Discussion of RQ$_3$.*

Our results reveal a disconnect between the perceived importance of replications and their acceptance. Despite the majority of participants indicating that they value replications, and that the field needs more of them, there is no shared understanding of when a replication is valuable or publishable. Establishing a common understanding and guidelines on what constitutes a valuable replication could enhance the conduct and publication of such studies.

Interestingly, participants also point to several social dynamics that might impact the likelihood of a researcher to conduct a replication. Beyond mistrust in same-author replications, one participant mentions that

> *"[..] you have to be really sure of yourself because it feels like you're attacking the other work if you find issues."*$_{P23}$.

This indicates an underlying assumption that, if the results of a study cannot be replicated, it is due to errors by the original authors or a clear indication that an effect does not exist. It seems that parts of the community are unaware that there are multiple reasons for failed replications. Especially in quantitative studies, mixed evidence (i.e., studies in favor of an effect and against it) is to be expected [18]. So, we should advocate for replications and expect them also to fail in a certain number of cases; analyzing these series of studies then might point us toward proving existence and better understanding of an effect. Moreover, replications are not only relevant if they show contradictions, but should strengthen trust in *existing* effects and theories and increase validity as part of a series [24].

Finally, despite several recommendations for incentives, changing the research culture cannot be solely the responsibility of researchers and the community. Novel approaches that do not hold their promises due to different contexts or unknown confounding factors should be a concern of funding agencies and we need political and organizational will to reward replications, which would improve the overall quality of empirical research in our field.

Key Results RQ$_3$

- Participants recognize the role of replications, but also acknowledge that it is difficult to get them published.
- Participants suggest ways of publishing and incentives for replications. However, there is no agreement how to handle replications.
- There is a sentiment that only failed replications are worth publishing and that they relate to errors in the original study. However, mixed evidence in series of replication studies should be expected.

4.4 RQ$_4$: How does the research community think about ecological validity?

4.4.1 *Quantitative results.*

Figure 3 shows answers to the closed questions regarding RQ$_4$. Overall, the results indicate that many participants identified a conflation of the concepts external and ecological validity. A majority thinks that more awareness should be paid to the distinction and that the research community should

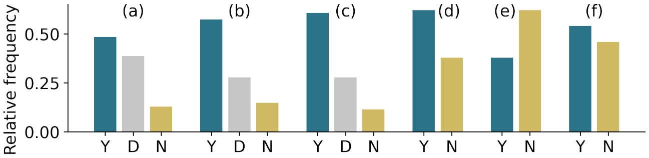

Figure 3: Closed questions for RQ$_4$: Y=Yes; D=I do not know; N=No for questions (a) Did you notice situations in which the concepts external and ecological validity were mixed up?; Do you think the software engineering research community should (b) ... be more aware of the distinction? (c) ... strive for more ecological validity?; If yes, which of the following would be acceptable to compromise on, when conducting more ecologically valid studies: (d) int. validity; (e) ext. validity; (f) replicability

strive for more ecological validity. However, there is no clear picture on what the community should compromise on when increasing ecological validity.

4.4.2 *Qualitative results.*

Overall, participants expressed a wide range of opinions on ecological validity, from *"LOVE IT"*$_{P68}$ to *"ignore it"*$_{P58}$. We will now dive into specific aspects of this discussion.

Distinction of external and ecological validity. Several participants have not heard of the term ecological validity before, some stated that they see external and ecological validity not as distinct concepts, and several state that they mixed up the two concepts themselves, also when answering the questions at the beginning of the questionnaire. Another participant describes that there is not only unawareness of ecological validity, but rather that the community generally is prone to several misunderstandings of validity concepts:

> *"In at least 90 % of papers I review, the limitations section demonstrates completely misunderstanding of common criteria. People don't know the difference between internal and external validity let alone ecological. NO ONE seems to know what construct validity means.[..]"*$_{P46}$.

Thus, while participants know and distinguish internal and external validity, the term ecological validity is less known.

Compromising for ecological validity. Most participants who think we need increased ecological validity indicated that there should be a reasonable trade off between internal, external, and ecological validity, as well as replicability. Some participants indicate that the compromise depends on several aspects, such as topic, methodology, or *"[..] the audience for the research."*$_{P30}$, that is, what to compromise on is not a generic decision, but depends on the context.

Reservation toward increasing ecological validity. Several participants express reservations toward increasing awareness and more emphasis on ecological validity, with some saying that the research community is not mature enough, or that the terminology generally does not matter:

> *"From a methodological standpoint, many researchers are not prepared enough to recognize the difference."*$_{P10}$,

> *"Coming up with more definitions and classifications is not helpful."*$_{P35}$.

In the same vein, one participant indicates that the notion of ecological validity is already used *"[..] as a blunt rejection instrument [..]"*$_{P6}$, although reviewers do not name the concept explicitly. Whether ecological validity should be emphasized more, depends on whom the research targets:

> *"[..] If the objective is to make SE research relevant to industry, then yes [emphasize more]. If the objective is to study the phenomena of software engineering practice, human cognition and knowledge organization, etc., then maybe not."*$_{P27}$.

This indicates that different criteria apply for different goals.

Ecological validity and the relationship to industry. Several participants indicate that ecological validity is for them closely related to industry-led studies and industry collaborations. While some more positively say that ecological validity is useful *"[..] for attracting collaborators from industry [..]"*$_{P22}$, others have severe reservations toward studies that are led by single companies:

> *"I'm increasingly uncomfortable with industry led studies. I do not take seriously papers on how great Microsoft tools are from Microsoft researchers. But then who does those studies, if Microsoft is not involved?"*$_{P25}$.

Thus, some questions might only be investigated through industry collaborations, but this way the relationship to independent research may be compromised.

4.4.3 *Discussion RQ$_4$.*

Overall, the research community is less aware of the term ecological validity, instead it refers simply to external validity. This corroborates our insights obtained for RQ$_2$ that there is a conflation of external validity and how realistic a setting is, or how practically applicable findings are. When confronted with a definition of ecological validity, most participants acknowledge this conflation.

While many participants advocate for a pragmatic approach with less emphasis on terminology, their responses reveal a wide range of design preferences and opinions: Consistent with results regarding the tradeoff between internal and external validity, most participants indicate a willingness to compromise on internal validity rather than on external validity. This finding is somewhat unexpected. If closeness to the real-world experience of software developers is the primary reason to favoring external validity, one would expect more willingness to compromise on external validity for it. This holds a certain potential for conflict, as especially highly ecologically valid studies (e.g., field studies) might not be generalizable [8].

Finally, participants indicate that ecological validity is often closely related to either industry collaboration or even industry-led studies, toward which participants express mixed feelings, including concerns about potential biases when studies are led by single (big) companies. This points to an underlying tension between the benefits of industry collaboration and the need to maintain unbiased, rigorous research.

Key Results RQ$_4$

- Participants recognize a conflation of external and ecological validity and indicate a need for clearer distinction and greater emphasis on ecological validity.
- Some participants express concerns about the community's maturity and toward industry-led studies.

4.5 Threats to Validity

Internal validity: A key threat is potential selection bias. Participants in this survey may be more interested in empirical research and thus exhibit greater awareness of methodological issues, find replication more or less important, or have completely different views than the general population of key players in the community. However, our results can be seen as representing the minimal spectrum of opinions within the community, which is already very diverse and illustrates one of the key points of this paper. Regression to the mean [3] additionally threatens the comparison of today's results with those from the original study. This statistical phenomenon can cause extreme values to appear less extreme when measured again, potentially skewing our comparison. However, our quantitative results do not indicate a change in the distribution of answers. As in the original study, the Rosenthal effect [29] may pose a threat to internal validity. That is, the way the questions are phrased might influence participants answers. In the original study, the way external validity was presented, implied that the goal is to "*create a practical, everyday setting.*", which might has caused the conflation with ecological validity. Interestingly, we also observe this conflation, despite having fixed this phrasing in our questionnaire.

External validity: Our survey investigated opinions of key players in the research community; it is unclear whether these insights generalize to all members of the software engineering research community, including authors who are not necessarily reflected in our sample. However, the original and the replication study were not designed to answer questions beyond the surveyed population. Thus, within the scope of gathering key players' perspectives, our results remain valid. Finally, we observed that some participants focused heavily on the example scenarios, potentially finding it difficult to abstract from them. As a result, their answers might be specific to the scenarios than reflecting broader concepts of validity, but still reflect the diverse opinions within the community.

Statistical conclusion validity: There is an unknown overlap of participants between our sample and the sample of the original study, which may violate the assumption of independence. We conducted a sensitivity analysis and found that such a violation would have no impact on the results.

5 Discussion and Perspectives

Having presented our results, we now compare our insights from the replication study to the findings from the original study and outline perspectives for future advancements.

5.1 What Has Changed in Ten Years?

We structure the discussion by RQ_1 to RQ_3, which were also part of the original study.

Overall, we see an increased awareness of the tradeoff between internal and external validity. Additionally, participants in the replication study show a richer understanding of different empirical methods and their answers are generally more nuanced. However, despite the ongoing discussions on guidelines and research quality, there is still a considerable disagreement among the reviewers regarding a multitude of views on how to approach and how to review (quantitative) empirical research.

5.1.1 *Awareness and unawareness of the tradoff.*

Ten years ago, results suggested a divided community. While there was evidence of some understanding about the tradeoff between internal and external validity, some participants made extreme statements indicating unawareness or ignorance. For instance, one participant stated that maximizing internal validity *"[w]ould show no value at all to SE community"*; we did not find such extreme statements in the sample of the replication study.

Furthermore, the discussion within the community on the quality of empirical research has clearly evolved. Participants in our study mentioned that different methods require different quality criteria and that internal and external validity might not apply to all research questions and designs. Even when explicitly searching for such a discussion in the raw data from 2014, we found only few participants who mentioned that they prefer qualitative or mixed methods, but no greater discussion. This broader sensitivity toward different methods observed today might be due to a diversification of empirical methods within the software engineering research community. Unfortunately, reviewing guidelines and standards do not reflect this diversification, yet. Promoting impact, novelty, and soundness without clear guidelines can lead to inconsistent evaluations of different research methods.

5.1.2 *Views on how to address the tradeoff.*

There are no significant differences on what kind of validity participants prefer today and ten years ago: It is still the case that the majority favors external over internal validity, with many stating that it depends on various aspects of the study, or that multiple studies should be conducted. Similarly, participants still hold strong and partly opposing views which population (e.g., students or developers) should be recruited.

When examining reasons why participants favor external validity, similarly to ten years ago, we observe that participants strongly associate external validity with how realistic a setting is. External validity is still seen as producing practically relevant insights for industry compared to internal validity.

5.1.3 *The role of replication.*

Participants still indicate that they value replications, that the community should emphasize them more, and that they are likely to accept more replications as their fellow reviewers.

The results of both studies show that participants value a certain delta in replications compared to the replicated study. However, it remains unclear whether this delta should aim at increasing internal or external validity, or addressing suboptimal design decisions of the original study. Furthermore, our results reveal that some believe only replications with contradictory results are valuable, while others view them as evidence of misconduct by the original authors. This indicates misconceptions about the purpose of replications, a sentiment not found in the data from ten years ago. Additionally, the indicated mistrust toward researchers who conduct replications of their own work, was not expressed as clearly ten years ago.

Overall, when we consider community efforts on improving the understanding of validity and empirical methodologies, such as the original study [37], books [14], standards [26], and other papers [4, 27, 28, 38, 41, 44], we see generally more educated reviewers, but still no structural and organizational changes on a large scale, such as reviewer assignments based on experience and expectations on empirical methods.

What has changed in 10 years?

- ► Participants in the replication study are more aware of the tradeoff between different kinds of validity.
- ► Participants still strongly associate external validity with realistic research settings and practical impact.
- ► Views on how to address the tradeoff between internal and external validity have not considerably changed.
- ► Participants still acknowledge the need for more replications, and the emphasis on novelty is still seen as a hindering factor.

5.2 Perspective and Call to Action

Having replicated the study from ten years ago, our insights lead to new conclusions and implications how to proceed as a community, which we discuss in detail next.

► *Distinction of different kinds of validity can help discussing fundamentally opposing positions within the community.* Participants in our replication survey find it generally important to differentiate between ecological and external validity. While we agree with some participants that the terminology itself generally does not matter too much, it still provides us with a reference framework to precisely talk about issues that the community is already concerned with. For instance, driven by the lack of industrial relevance in software engineering research, Briand et al. introduced a novel paradigm called *context-driven* research, which defines research problems according to industry needs [8]. By solving practically relevant problems with realistic assumptions and evaluations (ecologically valid research), software engineering research can bridge the gap between academia and industry. However, Briand et al. are willing to trade off the generalizability of research results: "*The fact that solutions and results don't generalize to all contexts [externally valid research] shouldn't be of concern.*" [8]. Prioritizing ecological validity over external validity is a valid approach, but we found counter arguments, such as a lack of independence of research and the inability to transfer insights from a large company to any other company, thereby questioning the value of the results. Instead of treating context-driven research as a separate paradigm, we suggest extending the existing framework of validities to include ecological validity, as this is also recognized and discussed in other fields.

► *Improved reviewing processes, guidelines, and empirical standards.* Our results show that reviewers might hold very different values and perspectives, despite reviewing the same paper. Given alone the large variety of views on ecological validity, from *"LOVE IT"*$_{P68}$ to *"ignore it"*$_{P58}$, our results show how large the spectrum of opinions is. As a consequence, a paper under review might be judged by a homogeneous group of reviewers not caring for this kind of study, a homogeneous group being too optimistic about a methodology, or a heterogeneous set of reviewers, whose decision might be dominated by a persistent reviewer. All cases could lead to biased judgments of the paper itself. This might contribute to a kind of randomness on whether a paper gets accepted or rejected, which has been shown to exist in review processes [19] and is a non-scientific assessment of the work, as expectations and not the quality of research matter. Reviewing guidelines provided to reviewers should accommodate different scenarios, ensuring that reviewer preferences do not overshadow or misalign with different works. Precise terminology in the papers and the reviewer profile and a clear understanding of implications and tradeoffs are essential.

Currently, reviewing guidelines often use terms, such as *quality*, *soundness*, or *rigor*, to refer to the quality of a study, leaving much room for interpretation by reviewers (despite more detailed explanations in the guidelines). Clearly, judging research quality is inherently challenging (which is why we conduct peer reviews), but there should still be a common ground for evaluating studies using different empirical methods. To help researchers identify valid and invalid criticism based on research methods, reviewing guidelines should foster and rely on standards for different aspects of research work, such as research questions, methodology, and context. For each of these, we should develop and maintain community standards, such as the ACM SIGSOFT Empirical Standards [26] for different methods. This would help reviewers and authors alike by allowing them to judge different research fairly and appropriately to the methodological specifics, and help write high quality manuscripts according to an agreed standard.

Finally, guidelines alone are not sufficient. The paper assignment process should also consider methodological preferences and epistemological beliefs, rather than solely focusing on a paper's topic. One way to account for more than just the topic are matching systems that match submissions to reviewers based on similarity between submissions and the reviewer's papers [9].

► *To replicate or not to replicate.* Most reviewers in our study find replications valuable, yet acknowledge that they are rarely published. A key reason is emphasizing on publishing *novel* research, which is also considered essential for a successful career. This contradicts the impression of a need for replications. If replications should become a standard in our field, we need a cultural change beyond the community's efforts: funding agencies and institutions must acknowledge the necessity of replications of certain projects.

However, this discussion has been ongoing for at least 25 years [5] with little change in the status quo of published replications. We should ask ourselves whether it is even possible for disciplines, such as computer science or software engineering, to develop an empirical standard on par with other sciences, such as medicine. Given the rapid pace of advancement in some areas, perhaps replications are not the right tool, because generated theories or explanation of phenomena might be outdated once or even before they are published. Our study cannot answer this general question, but we believe the community should continue this discussion, identifying fields where replications are indispensable and those where they are not worth the effort. This way, efforts to foster replications can be targeted and streamlined.

Acknowledgments

We thank all participants for their time and effort. This work has been supported by the German Research Foundation (DFG) under Grant SI 2171/3-2 (699378) and through Collaborative Research Center TRR 248 – CPEC (389792660), by the German Federal Ministry of Education and Research (BMBF, SCADS22B), by the Saxon State Ministry for Science, Culture and Tourism (SMWK) for the "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig" (ScaDS.AI), and by the European Union as part of ERC Advanced Grant "Brains On Code" (101052182).

References

[1] 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 (2019), 201–230.

[2] Klaus Backhaus, Bernd Erichson, Sonja Gensler, Rolf Weiber, and Thomas Weiber. 2021. Multivariate analysis. *Springer Books* 10 (2021), 978–3.

[3] Adrian G Barnett, Jolieke C van der Pols, and Annette J Dobson. 2004. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology* 34, 1 (08 2004), 215–220. https://doi.org/10.1093/ije/dyh299 arXiv:https://academic.oup.com/ije/article-pdf/34/1/215/1789489/dyh299.pdf

[4] Marvin Muñoz Barón, Marvin Wyrich, Daniel Graziotin, and Stefan Wagner. 2023. Evidence profiles for validity threats in program comprehension experiments. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1907–1919.

[5] Victor R Basili, Forrest Shull, and Filippo Lanubile. 1999. Building knowledge through families of experiments. *IEEE Transactions on Software Engineering* 25, 4 (1999), 456–473.

[6] Andrew Begel and Thomas Zimmermann. 2014. Analyze this! 145 questions for data scientists in software engineering. In *Proceedings of the 36th International Conference on Software Engineering* (Hyderabad, India) *(ICSE 2014)*. Association for Computing Machinery, New York, NY, USA, 12–23. https://doi.org/10.1145/2568225.2568233

[7] Brian H Bornstein. 1999. The ecological validity of jury simulations: Is the jury still out? *Law and human Behavior* 23 (1999), 75–91.

[8] Lionel Briand, Domenico Bianculli, Shiva Nejati, Fabrizio Pastore, and Mehrdad Sabetzadeh. 2017. The case for context-driven software engineering research: generalizability is overrated. *IEEE Software* 34, 5 (2017), 72–75.

[9] Laurent Charlin and Richard Zemel. 2013. The Toronto paper matching system: an automated paper-reviewer assignment system. (2013).

[10] Margarita Cruz, Beatriz Bernárdez, Amador Durán, Jose A Galindo, and Antonio Ruiz-Cortés. 2019. Replication of studies in empirical software engineering: A systematic mapping study, from 2013 to 2018. *IEEE Access* 8 (2019), 26773–26791.

[11] Daniela Damian and Andreas Zeller. 2021. ICSE 2022 Review Process and Guidelines. https://conf.researchr.org/getImage/icse-2022/orig/ICSE+2022+Review+Process+and+Guidelines-2.pdf.

[12] 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.

[13] Dror G Feitelson. 2022. Considerations and pitfalls for reducing threats to the validity of controlled experiments on code comprehension. *Empirical Software Engineering* 27, 6 (2022), 123.

[14] Michael Felderer and Guilherme Horta Travassos. 2020. *Contemporary empirical methods in software engineering*. Vol. 1286. Springer.

[15] Matthias Galster and Danny Weyns. 2023. Empirical research in software architecture—Perceptions of the community. *Journal of Systems and Software* 202 (2023), 111684.

[16] Gijs A Holleman, Ignace TC Hooge, Chantal Kemner, and Roy S Hessels. 2020. The 'real-world approach'and its problems: A critique of the term ecological validity. *Frontiers in Psychology* 11 (2020), 721.

[17] William Hudson. 2014. Card Sorting. In *The Encyclopedia of Human-Computer Interaction, 2nd Edition*, DA Bowman (Ed.). The Interaction Design Foundation: Aarhus, Denmark, Chapter 22.

[18] Daniël Lakens and Alexander J Etz. 2017. Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science* 8, 8 (2017), 875–881.

[19] Robert Langford. 2015. The NIPS Experiment. https://cacm.acm.org/blogcacm/the-nips-experiment/

[20] Jennifer R Ledford, Emilie Hall, Emily Conder, and Justin D Lane. 2016. Research for young children with autism spectrum disorders: Evidence of social and ecological validity. *Topics in Early Childhood Special Education* 35, 4 (2016), 223–233.

[21] Jenny T Liang, Chenyang Yang, and Brad A Myers. 2024. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *icse*. 1–13.

[22] Alina Mailach, Siegmund Janet, Sven Apel, and Norbert Siegmund. 2025. Supplementary Material to "Views on Internal and External Validity in Empirical Software Engineering: 10 Years Later and Beyond". https://doi.org/10.5281/zenodo.17095155

[23] Nasser Mustafa, Yvan Labiche, and Dave Towey. 2019. Mitigating threats to validity in empirical software engineering: A traceability case study. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2. IEEE, 324–329.

[24] Brian A Nosek and Timothy M Errington. 2020. What is replication? *PLoS biology* 18, 3 (2020), e3000691.

[25] Katherine Osborne-Crowley. 2020. Social cognition in the real world: reconnecting the study of social cognition with social reality. *Review of General Psychology* 24, 2 (2020), 144–158.

[26] Paul Ralph, Rashina Hoda, and Christoph Treude. 2020. ACM SIGSOFT Empirical Standards.

[27] Paul Ralph and Ewan Tempero. 2018. Construct validity in software engineering research and software metrics. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. 13–23.

[28] Martin P Robillard, Deeksha M Arya, Neil A Ernst, Jin LC Guo, Maxime Lamothe, Mathieu Nassif, Nicole Novielli, Alexander Serebrenik, Igor Steinmacher, and Klaas-Jan Stol. 2024. Communicating Study Design Trade-offs in Software Engineering. *ACM Transactions on Software Engineering and Methodology* (2024).

[29] Robert Rosenthal and Lenore Jacobson. 1966. Teachers' Expectancies: Determinants of Pupils' IQ Gains. *Psychological Reports* 19, 1 (1966), 115–118.

[30] Adrian Santos and Natalia Juristo. 2018. Comparing techniques for aggregating interrelated replications in software engineering. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. 1–10.

[31] Adrian Santos, Sira Vegas, Markku Oivo, and Natalia Juristo. 2019. A procedure and guidelines for analyzing groups of software engineering replications. *IEEE Transactions on Software Engineering* 47, 9 (2019), 1742–1763.

[32] Adrian Santos, Sira Vegas, Markku Oivo, and Natalia Juristo. 2021. Comparing the results of replications in software engineering. *Empirical Software Engineering* 26 (2021), 1–41.

[33] Simone G Shamay-Tsoory and Avi Mendelsohn. 2019. Real-life neuroscience: an ecological approach to brain and behavior research. *Perspectives on Psychological Science* 14, 5 (2019), 841–859.

[34] Martin Shepperd. 2018. Replication studies considered harmful. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*. 73–76.

[35] Martin Shepperd, Nemitari Ajienka, and Steve Counsell. 2018. The role and value of replication in empirical software engineering results. *Information and Software Technology* 99 (2018), 120–132.

[36] Janet Siegmund and Jana Schumann. 2015. Confounding parameters on program comprehension: a literature survey. *Empirical Software Engineering* 20 (2015), 1159–1192.

[37] Janet Siegmund, Norbert Siegmund, and Sven Apel. 2015. Views on internal and external validity in empirical software engineering. In *Proc. Int. Conf. on Software Engineering (ICSE)*. IEEE, 9–19.

[38] Dag IK Sjøberg and Gunnar R Bergersen. 2023. Improving the Reporting of Threats to Construct Validity. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*. 205–209.

[39] Klaas-Jan Stol and Brian Fitzgerald. 2018. The ABC of software engineering research. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 27, 3 (2018), 1–51.

[40] Margaret-Anne Storey, Neil A Ernst, Courtney Williams, and Eirini Kalliamvakou. 2020. The who, what, how of software engineering research: a socio-technical framework. *Empirical Software Engineering* 25 (2020), 4097–4129.

[41] Roberto Verdecchia, Emelie Engström, Patricia Lago, Per Runeson, and Qunying Song. 2023. Threats to validity in software engineering research: A critical reflection. *Information and Software Technology* 164 (2023), 107329.

[42] Nadia M Wager, Simon Goodson, and Loren E Parton. 2021. A systematic review of experimental studies investigating attitudes towards sexual revictimization: Findings, ecological validity, and scientific rigor. *Journal of criminal justice* 75 (2021), 101832.

[43] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2000. *Experimentation in Software Engineering: An Introduction*. Springer, Germany.

[44] 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*. 13–16.

[45] Zhiqing Zhong, Shilin He, Haoxuan Wang, Boxi Yu, Haowen Yang, and Pinjia He. 2025. An Empirical Study on Package-Level Deprecation in Python Ecosystem.

[46] Xin Zhou, Yuqin Jin, He Zhang, Shanshan Li, and Xin Huang. 2016. A map of threats to validity of systematic literature reviews in software engineering. In *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 153–160.

[47] Thomas Zimmermann. 2016. Card-sorting: From text to themes. In *Perspectives on data science for software engineering*. Elsevier, 137–141.