Threats to Validity in Software Engineering – hypocritical paper section or essential analysis?

Patricia Lago[*]
p.lago@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Per Runeson[*]
per.runeson@cs.lth.se
Lund University
Lund, Sweden

Qunying Song[*]
qunying.song@cs.lth.se
Lund University
Lund, Sweden

Roberto Verdecchia[*]
roberto.verdecchia@unifi.it
University of Florence
Florence, Italy

ABSTRACT

Background: In recent years, a discourse on how to systematically consider and report threats to validity started to gain momentum within the empirical software engineering community. **Aims:** With this study, we aim to systematically underpin the current state of threats to validity practices in software engineering research. **Method:** We conduct a literature review comprising 91 papers awarded with the ACM SIGSOFT Distinguished Paper Award at the ACM/IEEE International Conference on Software Engineering. Data is extracted and analyzed by considering six main facets of threats to validity, e.g., their explicit documentation, categorization, discussion of limitations, and trade-offs. **Results:** Results corroborate current critiques to the threats management state of the art. Threats result to be seldom discussed in depth, and are mostly considered as an enforced afterthought rather than an active concern of the research design and execution. **Conclusions:** To improve the observed practice, we derived items to consider for researchers, reviewers and readers, and call for a community action to increase the understanding of knowledge creation in empirical software engineering research.

CCS CONCEPTS

• **Software and its engineering**;

KEYWORDS

Empirical studies, Threats to validity, Limitations, Literature study

ACM Reference Format:
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 Proceedings of the 18th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '24).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3674805.3686691

1 INTRODUCTION

Empirical evaluations have become commonplace in software engineering research [28]. Demands on rigorous and transparent evaluations add value to both research and practice through assessment of current software engineering practice and proposed solutions to address identified problems. However, all empirical studies are not equally valid, or may be valid in different respects or contexts. Therefore, *threats to validity (TTV)* are discussed and analyzed to judge the validity of empirical studies.

In many research communities including software engineering [7], it is a recommended practice to discuss these in a TTV section with a twofold purpose:

i) provide a clear understanding of how the results are positioned within their context and what could have influenced the findings, and
ii) report mitigation strategies, i.e., how threats were alleviated, and/or why it was impossible to do so.

Other research traditions prefer the notion of *limitations*, which is a broader concept, referring to the choice of options that may cause TTV [24]. This is particularly common in design-science type research in computer science.

The presence and quality of TTV sections have been discussed in program committees and editorial boards. Concerns of weaknesses of TTV sections were raised more than a decade ago [5] and recently revisited in position papers [15, 24]. They have been criticized for being hypocritical sections of "laundry lists" of TTVs, rather than an effective contribution to the interpretation and use of the research results. Further, priorities or trade-offs between different categories of validity threats in relation to research goals are discussed conceptually [22], although not much in practice. However, the current status of TTV sections is not systematically explored, except for Sjøberg and Rye Bergerson's analysis of construct validity [20].

We therefore launched a study to assess current practices of TTV analysis and reporting in software engineering. Particularly, we wanted to study top ranked research, as an indication of what is considered high quality research. Our aim is to characterize how our community uses and communicates TTV. The observations range over a decade to enable identification of potential improvement trends in the community with increasing awareness and focus on TTV. Further, we aim to influence the community towards a more

[*] All authors contributed equally to this research.

ESEM '24, October 24–25, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-1047-6/24/10
https://doi.org/10.1145/3674805.3686691

transparent and useful TTV analysis and reporting, contributing to improved research quality.

Related work is summarized in Section 2 and our research approach – including TTV of this study – in Section 3. The results are presented in Section 4 and discussed in Section 5, resulting in twelve items to consider on TTV. Section 6 concludes the paper and outlines further work for our research community.

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.13382821, reference number [9].

2 RELATED WORK

Feldt and Magazinius reviewed 43 papers published in the ESEM conference in 2009. They analyzed the validity analysis in the papers to observe which threats and strategies for overcoming them that were reported [5]. They observed that 20% of the papers did not include any TTV analysis, which they call “an alarmingly high figure”. They also noted that established terminology was not used. To mitigate the problems observed, they proposed that a generic research process model be developed and used as a framework for TTV analysis. The same year, Wright et al. [26] stressed the need to focus on external validity by not only doing research on open source software, but also include proprietary software as study objects.

A few years later, similar to Feldt and Magazinius [5], Neto and Conte also proposed a conceptual model to solve TTV issues, although they model the relations between different categories of validity for controlled experiments, rather than the research process [13]. Siegmund et al. surveyed SE researchers, particularly program committee members, about their preference regarding the trade-off between internal and external validity [19]. They found that the opinions varied significantly and no consensus could be found on whether internal or external validity is most important for software engineering research. Petersen and Gencel analyzed how different TTV categories relate to different scientific world views [14] and recommend the researchers to report their worldview and analyze threats accordingly from a suitable categororization of TTV.

More recently, in a position paper, Verdecchia et al. [24] discuss whether TTV is reflected upon and an integral part of study design and analysis, or it is only reported post-study as a “laundry list” section. Robiliard et al. [15] similarly claim that 1) TTV sections are “boilerplate text by rote”, 2) encourage defensive writing style, and 3) is opaque about rationale for designs. Instead, they propose researchers to report *trade-offs* in research study designs, i.e. decisions points, alternatives, considerations, rationale for decision and implications.

Few singular research activities focused on specific categories of TTV, although recently some studies have contributed with solution proposals. Apostolos et al. [1] present a classification schema for reporting TTVs of *secondary studies*, and associated mitigation actions, are documented. The work of Apostolos et al. constitutes in our opinion a fresh outlook on managing TTVs, and an implicit acknowledgment of the overall suboptimal TTV consideration in current empirical software engineering literature.

Sjøberg and Rye Bergerson analysed the state of understanding and reporting practice of *construct validity* in empirical software engineering [20]. They analyzed articles in five SE journals, and observed an increasing appearance of construct validity analyses, although they also found a lack of adherence to established definitions. To support construct validity analyses, they defined a reference model and a set of guidelines to improve understanding and reporting of construct validity.

Some studies focus on sub-domains of software engineering. Malhotra and Khanna summarized TTV from 93 publications on search-based approaches for developing *software prediction models* [10]. They also summarize mitigation strategies against the threats. Mustafa et al. derived a sequence of validity analysis steps from research on *traceability* [12], which they advice to analyze and mitigate TTV. Wyrish and Apel summarize extensive work on addressing TTV in *program comprehension* studies [27]. They argue from their analysis that prioritizing TTVs is essential, and that this prioritization should be based on empirical evidence. Finally, Sanders et al. surveyed literature on *computing education research* [18] and aim to guide authors in addressing threats and limitations.

As outlined above, the interest for studies and guidelines have increased in recent years. However, the status of TTV analysis and presentations in general software engineering is not known.

3 RESEARCH APPROACH

The overall goal of our work is to:

- improve (purpose)
- TTV analysis and sections (issue)
- in software engineering research (object)
- from empirical software engineering researchers' point of view (viewpoint)

In this study, we explore the current practice in SE research to establish a foundation for future improvements. We explore which categories of threats are discussed, and whether authors distinguish between threats and limitations.

3.1 Research Method

We derived seven study questions and related metrics to categorize SE papers, as reported in Table 1. All metrics are categorial. Specifically, Q0 aims to identify the correlation between the type of study and threats and provides a context for other questions. Q1 assesses the quality of the TTV reflection in the selected papers. Q2–5 identify which checklists or guidelines and categories of threats are used. Q6 inspects whether authors distinguish between threats and limitations, and Q7 analyses if trade-offs in TTV are incorporated and reported. We also linked each question to one or more problems discussed in Verdecchia et al.'s position paper [24] to assess if there is empirical evidence for their claims.

As our pool of papers to study, we selected top ranked papers which are designated the ACM SIGSOFT Distinguished Paper Award[1] from the main technical track at the ACM/IEEE International Conference on Software Engineering (ICSE) across ten years of the conference (2014–2023, inclusive). In all there are 91 awarded papers. We chose ICSE because it is considered to be one of the top publishing venues in software engineering in several national and commercial ranking lists, that covers a broad set of diverse topics,

[1] https://www.sigsoft.org/awards/distinguishedPaperAward.html

QID	Question	Problem	MID	Metric
Q0	Which type of study is reported?		M0.1	Lab experiment/ Case study/ Mining study/ Interview study / Questionnaire study/ Literature review/ Simulation/ Design
Q1	Is TTV reflected upon generally in the paper?	P1, P5	M1.1	Dominantly "shallow" vs. "in-depth" reflection
Q2	Is a checklist or guideline used?	P1, P6	M2.1	Explicit reference to checklist/guideline?
			M2.2	Which checklist(s)/guideline(s) are referred to?
Q3	Which TTV categories are used?	P2	M3.1	Internal/external/construct/conclusion/reliability – credibility/transferability/dependability/confirmability/other
Q4	Is the TTV categorization fit for the study	P2	M4.1	Is the categorization used matching the current type of study?
Q5	Are there indications of proactive TTV analysis?	P4	M5.0	Is there a TTV (sub-)section?
			M5.1	Is TTV discussed in the research design/methodology section?
			M5.2	Are actions to mitigate TTVs discussed?
			M5.3	Is research design discussed in the TTV section?
Q6	Are limitations discussed?	P7	M6.1	Is there a (sub-)section entitled Limitations?
			M6.2	Are limitations discussed in line with established definitions?
Q7	Are trade-offs between TTV discussed?	P8	M7.1	Are trade-offs discussed in the research design/methodology section?
			M7.2	Are trade-offs discussed in the TTV section?

Table 1: Questions (Qn) and metrics (Mn.m) for the data collection. Problems (Px) refer to Verdecchia et al [24].

and we chose the "best" of those papers because we expected this cohort would represent exemplars of fine research. These might not be the "best" with respect to their TTV practices, but we wanted to analyze papers that are considered highest quality overall by the community. Random sampling of papers would give a general view of TTV practices in the community, while our approach focuses on what is considered top quality.

We created a data collection spreadsheet with metrics to fill out to answer questions for each paper. First, to calibrate our data collection, all four authors collected data for three selected papers from the latest year, representing three types of research (survey, interview study, and design). After calibrating the interpretation of the metrics scheme, each authors was assigned one fourth of the papers, every fourth paper on a temporally ordered list to mitigate the risk of bias in the analysis of the temporal development. The data collection was conducted in three batches of 5–10 papers per author, and borderline cases were discussed and agreed among the authors, for each batch. Then, descriptive analyses were conducted for each of the metrics. Consistency between the classifications was assessed during the analysis.

We analyzed the outcome of each metric with descriptive statistics. We also made a cross analysis of study types and TTV practice, as well as the development over time, by splitting the data set in two for the years 2014–2018 and 2019–2023, respectively. We did not conduct any statistical analyses since the data sets are small. Rather, we interpret the descriptive statistics, qualitatively.

Based on the findings, we analytically derived considerations for researchers, reviewers and readers at different stages of the research process, to mitigate the problems identified in the current status of TTV practices.

3.2 Threats to Validity Analysis and Mitigation

This study is a secondary study, although not a traditional systematic literature review or mapping study. We therefore divert from general TTV guidelines and consult specific ones for secondary studies in our analysis.

Kitchenham et al. recently proposed guidelines for reporting secondary studies, SEGRESS [8], based on standardization of secondary study reporting in medicine (PRISMA 2020[2]). They discard the notion of TTV in secondary studies, and introduce the term *Risk of Bias* (RoB) instead. It includes both RoB of the individual studies and the synthesis across studies. Furthermore, it uses *Certainty Assessment*, to assess the confidence in the body of evidence related to a specific finding.

We prefer to align with Ampatzoglou et al. [1], who based on a tertiary study in software engineering, proposed a TTV analysis schema for secondary studies, comprising study selection validity, data validity, and research validity. We discuss design considerations and trade-offs for our work accordingly below.

Study selection validity refers to search and filtering of primary studies. In our case, we have chosen to define a population of papers, based on them being published at ICSE and awarded ACM's distinguished paper award, thus avoiding search and filtering issues at all. Further, we have chosen a time span of a decade to enable observation of potential trends. A threat to study selection validity is that the papers accepted and awarded might not be the "best" papers in the SE community, but they represent peer review based decisions, anchored broadly in the SE community. Alternatively, we could have selected journal papers like Sjøberg and Rye Bergerson [20], but then we would not have an awarded subset. The core issue of our study selection is the concept of "best papers", which is not theoretically well defined. However, the selected set

[2] http://www.prisma-statement.org

of papers are awarded as "best" and hence represent some kind of best practice.

Data validity includes threats related to the data extraction and analysis steps of the study. To mitigate data validity threats, we calibrated our data collection process by classifying three same papers. Then we discussed cases which were considered borderline for each of the three iterations. We also assigned papers from all years to all authors, to avoid the risk of interpreting classification bias as trends – or lack of trends. We only use descriptive statistics and reason qualitatively about the results, since the data set would be too small for statistical analysis of the relevant correlations.

Research validity refers to the overall research process, such as chosen research method, generalizability, and repeatability. Given that the research goal is exploratory, rather than explanatory, we do not see the need to execute a full tertiary study to find indications of the state of TTV practice in SE research. Comprehensive literature reviews, as conducted by Ampatzoglou et al. [1] and Sjøberg and Rye Bergerson [20], would include the community in a broader sense, while we here focus on a specific subset of a specific conference, which is considered top ranked. The approaches thus complement each other. To promote repeatability, we calibrated our data collection process, and we make the raw data [9] openly available for others to analyze.

Regarding the validity of the considerations derived from the literature analysis (see Section 5.2), they are in the status of a proposal to the research community. We would appreciate further validation and extension of them within the software engineering research community.

4 RESULTS

4.1 Q0 – Which type of study is reported?

To investigate how threats to validity is being discussed in different types of studies, we extracted this information from each of the 91 selected papers. We discovered that 55 papers use a multi-method approach, i.e., the study they report combines two types of research methods. This is why the following presents both the total count of study types and, where applicable, which study types are composed in the selected papers. Naturally, in the first case the totals are more than 91.

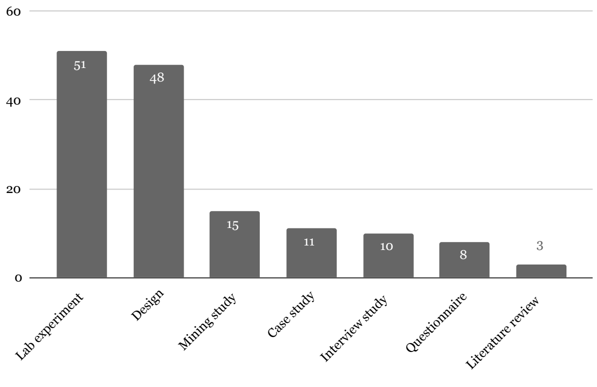

Figure 1: Count of covered study type(s) (Q0)

In particular, as illustrated in Figure 1, the selected papers report studies that are for the large majority designed as lab experiments (51 studies) and/or design studies (48). Further, we identified 15 mining studies, 11 case studies, 10 interview studies, 8 questionnaire surveys, and only 3 literature reviews.

As many papers report a multi-method type of study, it is further interesting to show which study types are being used within one study, either individually or in combination. In this respect, as illustrated in Figure 2, if we look at the top-5 study types used in the 91 selected papers, 34 papers (about 37.4%) combine design studies and lab experiments, followed by 14 papers (15.4%) lab experiments only, 6 papers case studies, 5 papers a combination of design study and mining study, and 5 papers design studies only.

Among the 55 papers (about 60.4%) reporting a multi-method type of study, in Figure 2 we see that the most-frequently used combination is Design and Lab experiment (34 papers). The second most-frequent combination is Mining study and Design (5) followed by Design and Case study (3) and Interview study and Questionnaire study (3). There are other combinations appearing in one or two papers each. Typically, combinations of study type involve a design type study and an empirical type study. The latter functioning as problem identification or design validation.

Finally, we further reflect that in the case of multi-method studies, one could expect that the related categories of TTV/limitations are blended in the common TTV discussion. While this is outside the scope of our study, it is also true that to the best of our knowledge, there is no systematic approach to provide such combination either. This could point to a possible future work, namely, studying if there is a correlation between the multi-method studies and the combined TTV, and eliciting possible guidelines.

While combinations of study types are extensively used in the selected papers, a systematic approach or guidelines for reporting TTV for such studies remains to be established.

4.2 Q1 – Is TTV reflected upon generally in the paper?

Figure 3 summarizes how many papers do include a reflection on the possible threats to the validity of their studies, and with what depths. Only 20 papers (22%, see label 'In-depth') discuss TTVs related to specific aspects of their study, while the majority (51 papers, about 56%, see label 'Shallow') dominantly relate general aspects of the used study type (e.g., case study) to generally applicable TTVs (e.g., generalizability). It is worth noticing that 20 papers do not report about the relevant TTVs at all.

Despite the primary studies being "best papers", most do address TTVs either shallowly or not at all. This further confirms the call for action as in Verdecchia et al. [24].

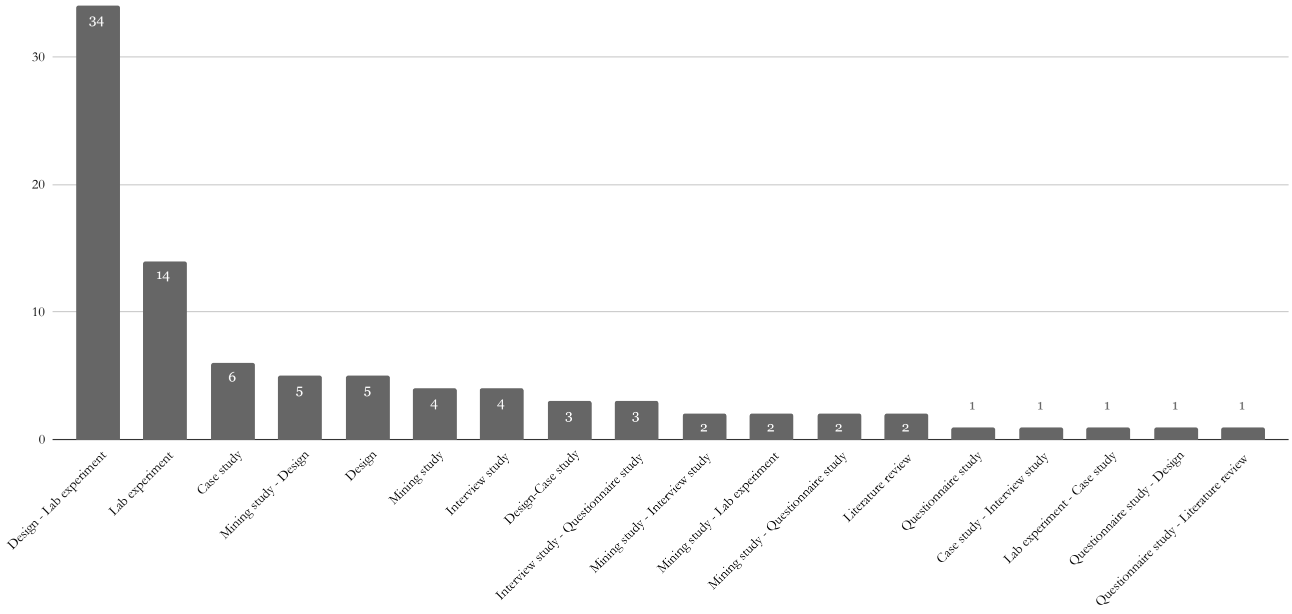

Figure 2: Compositions of study type(s) within same papers (Q0)

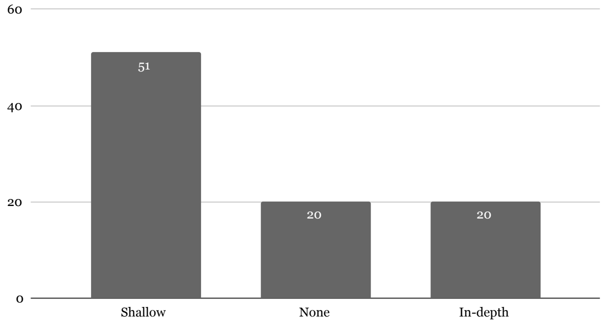

Figure 3: Depth of general reflection about TTV (Q1)

4.3 Q2 – Is a checklist or guideline[3] being used?

Only 2 out of 91 primary studies do explicitly refer to checklist or guideline with a categorization (and hence definition) of TTVs. In other words, most papers use terms but leave implicit where those terms come from, or how they are defined.

> Almost no studies explicitly refer to a checklist or guideline in the discussion of TTV, leaving why and how different categories of threats are adopted or defined unclear.

4.4 Q3 – Which categories of threats are being discussed?

A variety of categories of threats are used in the selected papers, suggesting no common guidelines are developed or generally accepted.

As shown in Figure 4, internal, external, and construct validity are the top three categories being discussed, accounting for 34 (37.4%), 29 (31.9%), and 19 (22%) papers, respectively. In contrast, conclusion validity and reliability are discussed in only 4 and 5 papers, respectively. Internal and external validity emerge from Campbell and Stanley [2], later extended by construct and conclusion, by Cook and Campbell [3], and conclusion replaced with reliability for qualitative studies [16]. In addition, 7 papers present either credibility, transferability, dependability, or confirmability in the discussion of TTV, which are categories used in qualitative research to assess trustworthiness [6].

Apart from those common categories selected, other categories are also used. Among them, 10 papers present generalizability (or generality and generalization), 3 on bias, and the rest include representativeness, reproducibility, verifiability, ecological validity, interpretive validity, convergent validity, internal consistency reliability, and discriminant validity. One may argue that some of them are similar or related to the common categories described earlier, however, we do not assume how the authors interpret or define them.

To better understand the presence of different categories, we analyse the overlapping between the papers that discuss the three most common categories of threats. As a result, 82.4% (28/34) of papers that discuss internal validity also present external validity threats. Conversely, 96.6% (28/29) of papers that discuss external validity appear to discuss internal validity. 50% (17/34) of papers

[3] We consider the two terms different, with checklist providing a descriptive list of tasks to be completed, and guideline suggesting a way how such tasks should be carried out.

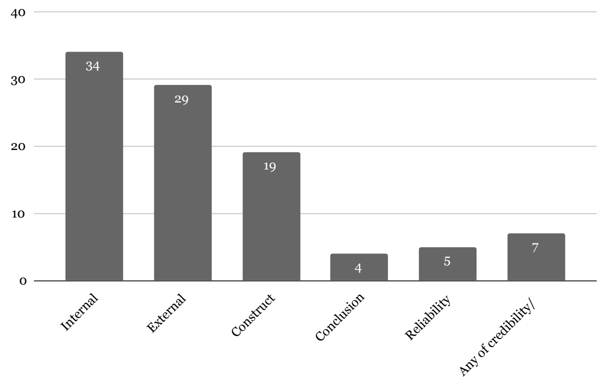

Figure 4: Categories of threats discussed in TTV (Q3).
The rightmost category is truncated due to space issues, and refers to "Any of credibility/transferability/dependability/confirmability".

that discuss internal validity also reflect on construct validity, and 89.5% (17/19) conversely. As visualized in Figure 5, there are very few (7/37, 18.9%) that discuss one of these categories exclusively. Thus, we find papers usually present several categories of validity in their discussion of TTV.

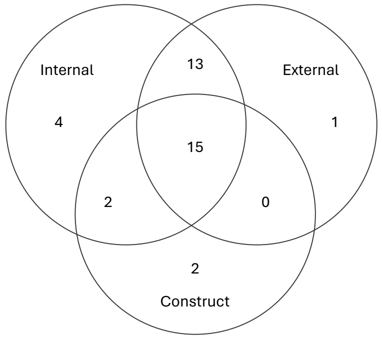

Figure 5: Overlaps of papers discussing internal, external, and construct validity and their threats (Q3)

Furthermore, we analyse the distribution of papers that discuss different categories of threats by time and depth of TTV discussion. As shown in Figure 6, we divide the papers into two 5-year intervals, and find the proportions of papers are almost even for each category of threat. That implies discussion of TTV and different threats do not change over time. In contrast, as shown in Figure 7, papers evaluated with in-depth reflection on TTV possess a significantly higher proportion than the shallow ones, especially in discussing internal, external, and construct validity.

Various categories of threats are used in the selected papers, with internal, external, and construct being discussed the most. Therefore, we believe a comprehensive solution to systematically identify, evaluate, mitigate, and report TTV is needed, considering different research goals and methods.

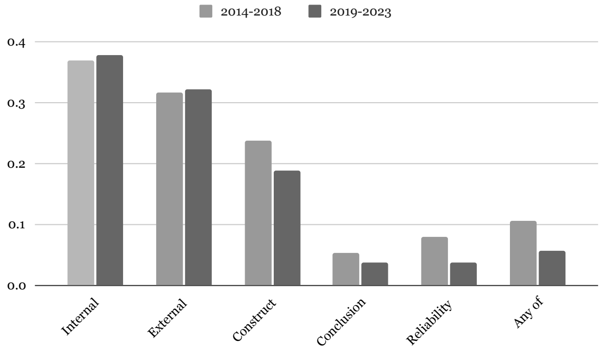

Figure 6: Proportions of papers in two 5-year intervals (Q3).
The rightmost category is truncated due to space issues, and refers to "Any of credibility/transferability/dependability/confirmability".

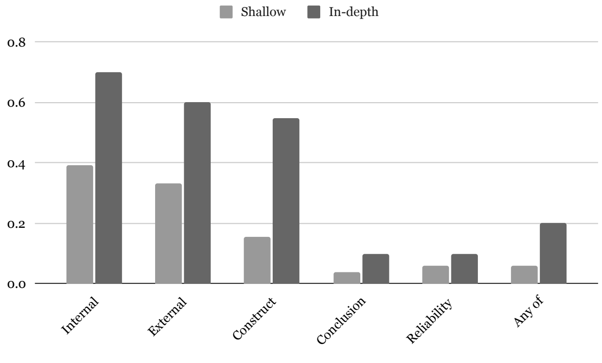

Figure 7: Proportions of papers by depth of TTV (Q3).
The rightmost category is truncated due to space issues, and refers to "Any of credibility/transferability/dependability/confirmability".

4.5 Q4 – Is the TTV categorization fit for the study?

Overall, there is a small amount (29/91, 31.9%) of papers that are considered using an appropriate categorization for analyzing and reflecting TTV. While those papers are almost evenly distributed in two 5-year intervals (i.e., 15 in 2014–2018, and 14 in 2019–2023), the proportions of them in corresponding intervals differ significantly, with 39.5% for the first and 26.4% for the second interval. Therefore, we observe a slight decline over time and recent papers do not perform better in adopting a fitting categorization for TTV discussion. The same distribution appears in another analysis of the same papers, with 15 labeled shallow on TTV reflection and 14 as in-depth. The respective proportions are 29.4% for the shallow papers and 70% for the in-depth ones. Given that, we observe a high correlation between the depth of the TTV reflection and the fitting categorization used.

Few studies use a suitable categorization for the study type in the discussion of TTV, suggesting no common guidelines are commonly accepted yet, thus; a comprehensive solution is urgently needed.

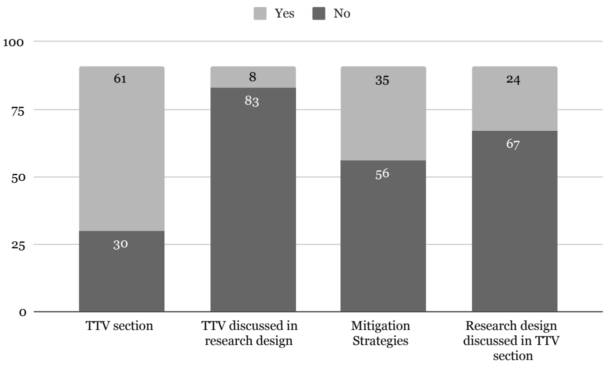

Figure 8: Indicators of proactive TTV analysis (Q5)

4.6 Q5 – Are there indications of proactive TTV analysis?

An overview of the recurrence of proactive TTV analysis indicators in the primary studies is documented in Figure 8.

As we can observe from the figure, while the majority of primary studies presents a section dedicated to discussing TTVs (61/91, 67.0%), a considerable portion does not (30/91, 33.0%). In the latter case, TTVs are either mentioned in another section (e.g., the Result or Discussion section), or not discussed at all (see Section 4.2). By considering exclusively the studies that discuss threats to validity, we observe that the recurrence of study types reflects the general one presented in Section 4.1. This suggests that no particular type of study is more likely to discuss TTV than the other ones.

By considering instead how often TTVs are discussed in the research design and/or methodology section, we note that this is done only in a minor portion of studies (8.8%).

A considerable number of primary studies does not report a discussion of mitigation strategies (56/91, 61.5%). As for the TTV section presence, no study type seems to be more prone to discuss mitigation strategies than the other ones.

Approximately a fourth of all primary studies discusses the research design in the TTV section (24/91, 26.4%). As could be intuitively expected, most papers discussing TTV in their research design correspond to studies that also report an in-depth TTV discussion (58.8% of papers reporting TTV in the design section discuss TTV in-depth).

> 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. Further, TTVs are almost never discussed in research design, which could be deemed as a consequence of considering TTVs only as an afterthought [24].

4.7 Q6 – Are limitations discussed?

Only a minor portion of primary studies present a section dedicated to discuss limitations (13/91, 14.3%). Most frequently, limitation appear in papers reporting a design or label experiment study type. Given the high recurrence of such study types however (see Section 4.1), we cannot conclude that such study types discuss limitations more frequently.

By considering how often limitations are discussed in line with our definition [24], the majority of papers reporting a limitation section results to do so (61/91, 67%). In addition, a small fraction of all papers (11/91, 12.1%) discussed limitations by adhering to our definition, while not reporting explicitly a limitation section.

While, from an optimistic point of view, this could be due to the absence of limitations in most of the selected studies, the overall picture drawn from this study appears to be more grim. Perhaps, to make sound contribution to the software engineering field in the future, the research community might need to take a step back from the current fast-paced publication-driven direction it might be embarking on. In doing so, limitations might be discussed with more ease, and study findings could be better understood, replicated, and reused.

> Limitations of studies are almost never reported and should be discussed appropriately to make sound contributions.

4.8 Q7 – Are trade-offs between TTV discussed?

Trade-offs between TTV are reported exclusively in two papers (2/91, 2.19%), which document the tradeoff, respectively, in the research design [4] and in the TTV section [23].

> Trade-offs between TTV is rarely discussed. Rather than a recurrent research documentation shortcoming, we could attribute this trend to a current lack of guidelines and education on studying, evaluating, and reporting TTV in the current empirical software engineering literature.

5 DISCUSSION

In this section, we report a discussion of the results and their potential implications (Section 5.1), accompanied by a set of considerations that can support researchers, readers, and reviewers in considering TTV (Section 5.2).

5.1 Findings and Implications

Based on our analysis of the selected papers, we observe that the discussion of TTV is dominantly shallow or absent, despite them being awarded the "best" papers in the top-ranked ICSE conference. Many papers do not use a fitting categorization or do not even possess a TTV section. Aside from that, TTV is almost never presented in the study design, and incorporates no tradeoff between different threats and their mitigation. In line with TTV, limitations are rarely discussed in the selected papers. Overall, our study reveals that current practices of TTV analysis and reporting in software engineering research are insufficient and require significant attention as well as improvements.

Our reflection, based on our empirical results, corroborates the position on TTV presented by Verdecchia et al. [24], which posits that researchers tend to take TTV as a checklist and afterthought, without considering a fitting TTV categorization, leading to a limited value that TTV sections seem to offer in current software engineering literature. Our results confirm the need for a systematic solution to address TTV in empirical software engineering, and, backed by empirical evidence, renew the call for action of Verdecchia et al. [24].

Moving a step forward, in the following we present further general reflections that go beyond the factual findings presented in Section 4.

By looking at the publications, we observe that there is **no improvement over time** in terms of TTV discussion coverage and depth. This in spite of the many initiatives carried out to improve the quality of design, execution, and report of empirical research in the software engineering field. One possible interpretation of this finding is that we focus too much on a *standard* structure of research documentation, and too little on the ultimate goal, namely, to openly document the limitations and validity of a study so that it can be replicated, extended, or transferred in different contexts by other researchers without encountering known issues, should they wish to do so.

The above would be in the interest of research. However, we also understand that if one would be *too honest*, they would also run a higher **risk to attract easy criticism** and ultimately being rejected by not-so-constructive reviewers. And then we argue: what are the incentives to be honest and discuss limitations and threats to validity in depth? Further, an extensive TTV analysis takes space in the paper, which authors may want to use to elaborate their contribution.

A less negative but still interesting interpretation, could be that standards like for example the ACM SIGSOFT Empirical Standards for Software Engineering, the many published guidelines, or even excellent structures of published empirical studies, are too often adopted at face value, **without full understanding**, critical thinking or further analysis. I.e., as already mentioned, as a checklist or afterthought. To counteract this trend, how can we learn and teach such deeper understanding and hence use checklists and guidelines for a useful TTV analysis?

The above reflections indicate that problems arise in the research community at large. Thus, potential countermeasures against the observed shortcomings cannot fall on the individual researchers. On the contrary, it must be a **community effort** to initiate improvement initiatives to ensure transparency about validity and limitations of research.

There is no lack of standards, guidelines and structures, although they are rarely referred to in the studied papers. What is missing is rather the in-depth understanding of the knowledge creation in software engineering and its scientific foundations. Based on that understanding, checklists and practices can be useful tools for insightful reasoning about research validity.

Literature for such development is also readily available. Storey et al. [22] and Stol and Fitzgerald [21] have provided solid foundations, by adapting the *generalizabilty–precision–realism* framework by McGrath et al [11, 17], to enable discussion and positioning of research validity in relation to research goals and methodology used. Robiliard et al [15] provide assistance for trading categories of validity threats. And the call for papers for ESEM already refers to these, requiring that "[p]apers should be positioned in terms of research methodology and contribution in relation to established frameworks".

5.2 Considerations for Addressing, Documenting, and Reviewing TTV

Based on our empirically-derived observations and resulting discussion, we propose a list of considerations, in the form of questions and suggestions, that empirical software engineering researchers may reflect upon when considering TTVs. An overview of the considerations is summarily reported for reference in Table 2. The considerations reported are not exclusively tailored for *researchers* who are conducting an empirical study, but also for *reviewers* who are scrutinizing a paper, and *readers* who are studying it and using the findings. The considerations are primarily intend to provide a perspectives to take into account for the three groups of stakeholders of a research study, as a step forward to improve current practices of TTV. They are separated into five phases of a research study, *design, implementation, analyses and reporting, review*, and *reading*.

Each consideration reported in Table 2 is mapped for convenience to an identifier (CID), its intended stakeholder, and the question of this research mapped to it (QID). In the following, each consideration presented in Table 2 is presented through a more comprehensive description, supporting rationale, and mapping to the findings of our literature review.

Design

C1 *Which categories of validity are most important for the given research goal?* As various categories of validity may emerge in a study, researchers should consider which category(ies) is most significant with respect to the research goals. This would be the starting point for a focused section reporting an in-depth discussion of the most relevant threats specific to the study at hand, moving away from a shallow threat analysis (Q1).

C2 *How are TTVs traded against each other?* As observed in the results of Q7, we noted that only two papers incorporated trade-offs between TTV. During the study design phase, researchers should focus not only on addressing TTVs, but also actively reason on how tradeoffs between threats may influence the validity of their study. This would allow to systematically evaluate and determine how different categories of validity should be traded against each other [15] and consciously prioritize the most relevant ones.

C3 *Which research method(s) are most suitable, given that goal?* Given the research goals a study aims to achieve, different research methods or a combination of multiple methods may be employed (see results Q0, Section 4.1). This practice will, however, expose studies to different threats to validity, and therefore requires targeted strategies to mitigate them appropriately. An exemplary reference for targeted threats are the TTV specific to systematic literature reviews presented by Ampatzoglou [1]. As different research methods may suffer from different types of threats, researchers should consider which research method(s) are most suitable to implement

CID	Phase	Question / Suggestion	Stakeholders	QIDs
C1	Design	Which categories of validity is most important for the given research goal?	Researchers	Q1
C2	Design	How are TTVs traded against each other?	Researchers	Q7
C3	Design	Which research method(s) are most suitable, given that goal?	Researchers	Q0
C4	Design	Which categories of TTV (or limitations) are most suitable for the chosen type of study?	Researchers	Q3, Q4
C5	Design	Is a suitable checklist or other guidance available for the type of TTV?	Researchers	Q2
C6	Implementation	Which choices are made in the implementation of the research design, that may impact TTV?	Researchers	Q7
C7	Implementation	Are there any established, evidence-based practices within the sub-field that can be used?	Researchers	Q2
C8	Analysis and reporting	Analyze TTV in depth.	Researchers	Q1
C9	Analysis and reporting	Present TTV and limitations in a balanced way. Write for the user of the research, and not for the reviewer.	Researchers	Q5, Q6
C10	Review	Assess TTV of study in relation to its contributions/claims.	Reviewers & Editors	Q0–Q7
C11	Review	Do not punish honesty in TTV/Limitations analysis.	Reviewers & Editors	Q0–Q7
C12	Reading	Assess/cite/use the results in relation to the TTV reporting.	Audiences	Q0–Q7

Table 2: Considerations (indexed by CID) for different stakeholders at different phases of a research study.

and maximize not only the outcome, but also the validity of the study [21, 22].

C4 *Which categories of TTV (or limitations) are most suitable for the chosen type of study?* After research methods are selected in the design of a study, potential threats to validity have to be identified. As articulated by Verdecchia et al. [24], and also observed in Q3 and Q4, many researchers only report common categories of validity, while not considering the fitness of the threat category in their study. The consequences are that TTV discussed being irrelevant, and not providing a fair assessment, as well as potentially relevant TTV not being addressed or reported. To change that, an essential question that researchers have to consider, during the design of a study, is which categories of TTV (or limitations) are most suitable for the chosen type of study.

C5 *Is a suitable checklist or other guidance available for the type of TTV?* Another consideration for researchers in the design phase, as a follow-up step to C3, is to consider whether a suitable checklist or guidance is available for the type of TTV identified. Verdecchia et al. [24] have argued that researchers tend to blindly take common checklists for discussing and reporting TTV in their studies, without considering the fitness and completeness. Our results in this study, and more specifically the ones of Q2, somewhat mirror this position by identifying only two papers that cited a TTV checklist or framework explicitly. As an alternative, researchers should explore available resources, e.g., Wohlin et al. [25] for controlled experiments, and Ampatzoglou et al. [1] for secondary studies, identify fitting checklists, and look for guidance in handling TTV for their studies. This would allow researchers not only to more systematically focus on the TTV specific to their research, but also to discover TTV characteristic of their research they might have been unaware of.

Implementation

C6 *Which choices are made in the implementation of the research design, that may impact TTV?* TTV should be considered throughout a research study [24], rather than as after-thoughts. During the implementation of a study, different choices can be made for varying purposes or due to certain constraints. Essentially, researchers have to evaluate every choice they adopt and its potential impact on TTV. Similar to what we already articulate for C2 and C3, the choices must be made to conciously maximize the validity and balance various TTV of the study, as analysed in Q7.

C7 *Are there any established, evidence-based practices within the sub-field that can be used?* Although a comprehensive and universally accepted guideline for TTV may not exist, practices for a sub-field might have emerged for TTV when it comes to the implementation of a study, such as Wyrish and Apel [27]. It is important for researchers, when carrying out a study, to explore any established, evidence-based practices within the field of the study. Further, researchers should consider which practices are relevant and feasible to adopt, considering the method and actual implementation, if such practices exist. This would discourage researchers to reinvent already known mitigation strategies, and improve the systematicity and validity of their study through the adoption of already utilized and documented practices.

Analysis and reporting

C8 *Analyze TTV in depth.* After the implementation study phase, in the analysis and reporting phase, researchers should analyse and document TTV in depth (Q1). An in depth analysis of threats to validity includes several perspectives to reflect upon, such as (i) which categories of validity are considered the most important, (ii) which threats are identified, (iii) how are threats prioritized and mitigated, (iv) what adaptations due to threats have been included in the design, method, implementation, and analysis of the study, (v) what was the rationale which led to the selection of a certain mitigation, or finally, (iv) why it was not possible to mitigate a documented TTV. As additional consideration regarding the in depth analysis of TTVs, researchers need to consider the fitness of different categories or checklists available, and present a fair and honest assessment of TTV in the paper. As observed in the results of our study (Q0–Q7, see Sections 4.1–4.8) the selected papers, despite being selected as best papers in the ICSE conference, reported dominantly shallow or none TTV analysis, which clearly indicates that a complete analysis and fair assessment of TTV is not provided by most researchers.

C9 *Present TTV and limitations in a balanced way. Write for the user of the research, and not for the reviewer.* Similar to C7, researchers also need to balance the TTV and limitations of the study in their analysis and reporting, so that the audience can get a sufficiently accurate understanding of the TTV (Q5 and Q6, see Section 4.6–4.7). Being superficial in the analysis of TTVs or their documentation diminishes the quality of the study and, while it could be peer-reviewed and in the end published, would convey a very limited, or no, useful information to the audience of the paper. Keep in mind that the TTV is written for the audience, not the reviewers!

Review

C10 *Assess TTV of study in relation to its contributions/claims.* In the review phase of a research study, the responsibilities shift to the reviewers of a publication venue, who we believe should assess the TTV in relation to its contributions or claims made in the study. Such assessment should be anchored on the accuracy and consistency of the TTV analysis with respect to the contributions claimed, rather than simply checking the existence of a TTV section or the existence of common categories or checklists of TTV. In an optimal case, reviewers should provide a fair assessment of and useful feedback on the TTV, and in the worst case, reviewers may not even pay enough attention to the TTV analysis, which is definitely a situation that should be avoided. As discussed by Stol and Fitzgerald [21], and Storey et al. [22], it is paramount for reviewers to carefully interpret the threats and limitations of a study by actively considering the research method is uses, its goal, results, and conclusive claims. Again, the selected papers in our study reported dominantly shallow or none TTV analysis, which indicates that a fair assessment of TTV is not given by the reviewers.

C11 *Do not punish honesty in TTV/limitations analysis.* Reviewers should not punish honesty in TTV/limitations analysis. The focus remains on the contributions and validity of the study [21, 22], despite certain TTV may arise and be tolerated for varying reasons. As we emphasize in C8, researchers should provide an accurate and sound analysis of TTV. Reviewers, in turn, should not criticize an honest but fair TTV discussion in the paper.

Reading

C12 *Assess/cite/use the results in relation to the TTV reporting.* Finally, when a paper is published and disseminated to its potential audience, the audience should assess/cite/use the results in relation to its TTV reporting. That being said, given the contributions, relevance, significance, and novelty claimed in a study, one should still evaluate the validity of it with respect to its design, implementation, analysis, and reporting of the study. Only with enough attention from the community, the quality of TTV will be improved collectively in the future.

6 CONCLUSIONS AND FUTURE WORK

We follow up on recent discussions about TTV sections in software engineering papers [1, 24] by systematically assessing the current status of awarded papers in the ICSE conference, as considered top of the top by many academics in software engineering. Our findings corroborate earlier criticism of absent reflection and pro-active analysis of research validity. Our study shows no improvement trend over the decade of publications under study (2014–2023).

Available standards and guidelines for researchers are at best used as "laundry lists" and "boiler plates". Thus our research community must take initiatives to raise the level of understanding – among researchers, reviewers, and readers – of knowledge creation in empirical research, including transparent and humble discussions about validity and limitations of research contributions. We hope that our considerations will support such development in the community. Only by increased understanding of the who, what and why of software engineering research [21, 22], the TTV sections can be turned from being hypocritical afterthoughts to becoming essential for the design, analysis and interpretation of empirical software engineering research.

As future research direction, building on the gathered results, we envision to move beyond observing the TTV state of the art and explore how we can improve it. Our ultimate goal and open call for interested researchers is to understand how TTVs can be systematically addressed throughout all research phases, starting from the study design, to the execution, reporting, and reviewing processes.

ACKNOWLEDGMENTS

This work was supported in part by the Wallenberg AI, Autonomous Systems and Software Program (WASP); and by the Digital Sustainability Center (DiSC) at Vrije Universiteit Amsterdam. The authors thank anonymous reviewers for their constructive feedback that helped improve the manuscript.

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. https://doi.org/10.1016/j.infsof.2018.10.006

[2] Donald T. Campbell and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin Company, Boston, MA, USA.

[3] Thomas D. Cook and Donald T. Campbell. 1979. *Quasi-Experimentation – Design and Analysis Issues for Field Settings.* Houghton Mifflin Company, Boston, MA, USA.

[4] Edson Dias, Paulo Meirelles, Fernando Castor, Igor Steinmacher, Igor Wiese, and Gustavo Pinto. 2021. What makes a great maintainer of open source projects?. In *IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, virtual, 982–994. https://doi.org/10.1109/ICSE43902.2021.00093

[5] Robert Feldt and Ana Magazinius. 2010. Validity Threats in Empirical Software Engineering Research – An Initial Survey. In *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE'2010)*. Knowledge Systems Institute Graduate School, 374–379.

[6] Egon G. Guba. 1981. Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology* 29, 2 (1981), 75–91. https://doi.org/10.1007/bf02766777

[7] Andreas Jedlitschka and Dietmar Pfahl. 2005. Reporting guidelines for controlled experiments in software engineering. In *International Symposium on Empirical Software Engineering* (Noosa Heads, QLD, Australia). IEEE, 10 pp.–. https://doi.org/10.1109/ISESE.2005.1541818

[8] Barbara Kitchenham, Lech Madeyski, and David Budgen. 2023. SEGRESS: Software Engineering Guidelines for REporting Secondary Studies. *IEEE Transactions on Software Engineering* 49, 3 (2023), 1273–1298. https://doi.org/10.1109/tse.2022.3174092

[9] Patricia Lago, Per Runeson, Qunying Song, and Roberto Verdecchia. 2024. *TTV Data Extraction*. https://doi.org/10.5281/zenodo.13382821

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

[11] Joseph E. McGrath. 1995. Methology Matters: Doing Research in the Behavioral and Social Sciences. In *Readings in Human–Computer Interaction*, Ronald M. Baecker, Jonathan Grudin, William A.S. Buxton, and Saul Greenberg (Eds.). Morgan Kaufmann, 152–169. https://doi.org/10.1016/B978-0-08-051574-8.50019-4

[12] Nasser Mustafa, Yvan Labiche, and Dave Towey. 2019. Mitigating Threats to Validity in Empirical Software Engineering: A Traceability Case Study. In *IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2. 324–329. https://doi.org/10.1109/COMPSAC.2019.10227

[13] Amadeu Anderlin Neto and Tayana Conte. 2013. A conceptual model to address threats to validity in controlled experiments. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering* (Porto de Galinhas, Brazil) *(EASE '13)*. ACM, New York, NY, USA, 82–85. https://doi.org/10.1145/2460999.2461011

[14] Kai Petersen and Cigdem Gencel. 2013. Worldviews, research methods, and their relationship to validity in empirical software engineering research. In *joint conference of the 23rd international workshop on software measurement and the 8th international conference on software process and product measurement*. IEEE, Ankara, Turkey, 81–89. https://doi.org/10.1109/IWSM-Mensura.2013.22

[15] Martin P. Robillard, Deeksha M. Arya, Neil A. Ernst, Jin L. C. 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* 33, 5 (2024), 1–10. https://doi.org/10.1145/3649598

[16] Per Runeson, Martin Höst, Austen Rainer, and Björn Regnell. 2012. *Case Study Research in Software Engineering – Guidelines and Examples*. Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9781118181034

[17] Philip J. Runkel and Joseph E. McGrath. 1972. *Research on Human Behavior: A systematic Guide to Method*. Holt, Rinehart and Winston, Inc., New York.

[18] Kate Sanders, Jan Vahrenhold, and Robert McCartney. 2023. How Do Computing Education Researchers Talk About Threats and Limitations?. In *Proceedings ACM Conference on International Computing Education Research - Volume 1 (ICER '23)*. ACM, New York, NY, USA, 381–396. https://doi.org/10.1145/3568813.3600114

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

[20] Dag I. K. Sjøberg and Gunnar Rye Bergersen. 2023. Construct Validity in Software Engineering. *IEEE Transactions on Software Engineering* 49, 3 (2023), 1374–1396. https://doi.org/10.1109/tse.2022.3176725

[21] Klaas-Jan Stol and Brian Fitzgerald. 2018. The ABC of Software Engineering Research. *ACM Transactions on Software Engingeering and Methodology* 27, 3 (2018), 1–51. https://doi.org/10.1145/3241743

[22] 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, 5 (2020), 4097–4129. https://doi.org/10.1007/s10664-020-09858-z

[23] Michele Tufano, Fabio Palomba, Gabriele Bavota, Rocco Oliveto, Massimiliano Di Penta, Andrea De Lucia, and Denys Poshyvanyk. 2015. When and why your code starts to smell bad. In *IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, Florence, Italy, 403–414. https://doi.org/10.1109/ICSE.2015.59

[24] 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. https://doi.org/10.1016/j.infsof.2023.107329

[25] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-29044-2

[26] Hyrum K. Wright, Miryung Kim, and Dewayne E. Perry. 2010. Validity concerns in software engineering research. In *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research* (Santa Fe, New Mexico, USA) *(FoSER '10)*. ACM, New York, NY, USA, 411–414. https://doi.org/10.1145/1882362.1882446

[27] 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)*. ACM, 13–16. https://doi.org/10.1145/3643664.3648203

[28] Li Zhang, Jia-Hao Tian, Jing Jiang, Yi-Jun Liu, Meng-Yuan Pu, and Tao Yue. 2018. Empirical Research in Software Engineering — A Literature Survey. *Journal of Computer Science and Technology* 33, 5 (2018), 876–899. https://doi.org/10.1007/s11390-018-1864-x