Threats to Validity in Software Engineering Research: A Critical Reflection

Roberto Verdecchia[a], Emelie Engström[b], Patricia Lago[c], Per Runeson[b] and Qunying Song[b]

[a] University of Florence, Italy
[b] Lund University, Sweden
[c] Vrije Universiteit Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Threats to Validity, Software Engineering, Empirical Research

ABSTRACT

Context: In the contemporary body of software engineering literature, some recurrent shortcomings characterize how threats to validity (TTV) are considered in studies.
Objective: With this position paper, we aim to open a discourse on the current use of TTV sections. The goal of our position is to jointly reflect and systematically improve how we, as a research community, consider TTV in our studies.
Method: Based on our personal experience as researchers, authors, reviewers, and editors, we critically reflect on the treatment of TTV in current empirical software engineering literature.
Results: We discuss the key shortcomings of TTV consideration, including the failure to acknowledge different types of validity categorizations and the tendency to treat threats just as an afterthought. For each identified problem, we propose a vision for an improved state, intending to catalyze thoughtful engagement and improvements the way our community addresses TTV.
Conclusion: We posit there is an urgent need to reconsider how we approach, document, and evaluate TTV in software engineering research.

1. Introduction

Empirical software engineering (ESE) is a field of software engineering research (SE) that utilizes empirical methods to investigate software development processes, products, and their related phenomena. As with any empirical research, ESE studies are vulnerable to various threats to validity (TTV) that can undermine the reliability and validity of their findings. These threats can occur at different stages of the research process, such as study design, data collection, and analysis.

To document and discuss TTV, related sections are commonly included in ESE papers. The purpose of TTV sections is twofold: (i) provide a clear understanding of how the results are positioned within their context and what could have influenced the findings; and (ii) reporting mitigation strategies, *i.e.,* how threats were alleviated, and/or why it was impossible to do so.

Laudable effort was spent in TTV to openly document potential shortcomings. In the contemporary body of ESE literature, however, we note that some recurrent shortcomings still appear to be present in TTV sections. In many ESE studies, TTV often seem to be unfocused and treated in a rather superficial manner. Frequently, TTV sections seem to be included just as a mandatory component, rather than a critical reflection on the potential threats of studies.

Despite words of warning raised by earlier research [3, 6] and more recent secondary studies [1, 7], to date, TTV sections seem to be mostly formulated as "laundry-lists" of potential threats, lacking a thorough contextualization in the specifics of the study at hand. In our opinion, this approach undermines the value of TTV sections as, even if apparently systematic, they fail to provide a thorough assessment of the potential limitations of studies. Furthermore, the lack of a clear focus and superficial treatment of TTV can impede the clear understanding of the reliability of findings, leading to several problems related to the use and replication of study results.

To address this issue, it is crucial to consider TTV as an essential part of the empirical research process, rather than just a perfunctory requirement. Researchers should consciously take time to critically reflect on the TTV of their studies throughout all research phases, without blindly relying in a check-list fashion on diktat imposed by pre-existing TTV categorizations.

In the following, based on our personal experience, we critically reflect on the shortcomings we note in the TTV of current ESE studies and sketch our vision on how TTV sections could be improved.

2. Problem Statement: Current (mis)Use of Threats to Validity Sections

Using threats to validity as checklists hinders reflection **(P_1).** In TTV sections of SE research, a clear picture emerges. At first glance, most TTV sections might appear as thorough and systematic. By reviewing the sections with more care, however, we note that often a long "laundry-list" of TTV is shallowly discussed. The lack of focus, and the will to cover all types of threats, regardless of their relevance, often leaves little room to discuss in depth the threats which most matter in the specific research presented. In other words, the intention to cover a vast range of threat types for the sake of completeness and systematicity often dilutes the discussion regarding most important threats, *i.e.,* the ones

roberto.verdecchia@unifi.it (R. Verdecchia);
emelie.engstrom@cs.lth.se (E. Engström); p.lago@vu.nl (P. Lago);
per.runeson@cs.lth.se (P. Runeson); qunying.song@cs.lth.se (Q. Song)

which are most important to fully understand the context of the study, its applicability, replicability, and results validity.

The frequent dilution of relevant threats with trivial ones in current literature might originate from the unreflected use of pre-defined, *de facto* standardized, TTV categorizations, *e.g.,* the one emerging from general experimentation in the seminal book by Wohlin *et al.* [9]. Rather than making a pondered decision on which TTV are most relevant to be discussed for the study at hand, TTV often seem to be summarily considered without following a proper threat identification, analysis, and discussion. By using TTV categorizations as mere checklists, *i.e.,* boxes to be ticked with brief discussions, TTV sections tend to loose their semantic meaning, becoming an *academic exercise* contributing little to the quality of a work.

Using a specific threats to validity categorization can be misleading **(P_2).** As an additional reflection on current TTV discussions in SE literature, we note that TTV sections often fail to acknowledge that different types of TTV categorizations exist. In the literature, TTV categorizations and their corresponding threat types vary depending on the research method, philosophical stance, and abstraction level considered [5]. For example, the threats outlined by Wohlin *et al.* [9] are fitted to discuss the TTV of a controlled experiment, but only marginally apply to systematic literature review studies, for which other types of threats are more appropriate [1]. Similarly, quantitative and qualitative research generally face different threats [4]. Failing to recognize the existence of different TTV types, and different categorizations of them, frequently leads to blindly following the checklists authors are more accustomed to, without questioning if the checklist is the best fitted for a study, or if relevant threats not listed in the checklist might have influenced the research. Furthermore, if mixed methods are used or multiple investigations are reported together, different TTV checklists might need to be consulted.

Threats to validity pitfalls seem embedded in the SE research community **(P_3).** As an anecdote from our personal experience, the dogmatic trend of blindly selecting a TTV categorization and following it as a checklist might not stem exclusively from the authors of SE papers, but also from the peers who review them. The treatment of certain TTV categorizations as standard, syntactic, checklists might have become a practice so embedded in the SE research community that not only it is not questioned anymore, but is nowadays even enforced by peer-reviewers. This leads at times to rather alarming review comments, suggesting to discuss unrelated / irrelevant threats, or even using a TTV categorization that does not fit the research method adopted in a study.

Threats to validity are often just an afterthought **(P_4).** If we zoom out of TTV sections and consider the bigger picture, namely how TTV are positioned within the execution of SE studies, we can often observe another point of critical reflection. Rather than being considered from the start of a research, *i.e.,* the research design, TTV frequently appear to be considered *a posteriori*, an afterthought to be discussed when the research execution is over and the final documentation of the study is drafted. Failing to consider TTV and their mitigation strategies during the design and execution of the study may lead down a perilous path. By postponing considerations on TTV till the end of a study, it might be impossible to accurately analyze TTV and consciously design appropriate mitigation strategies.

Threats to validity sections fail to fulfill their true purpose **(P_5).** Overall, by considering our reflections on TTV sections in contemporary SE research papers, it becomes natural to ask ourselves the question: *"What is the purpose of writing threats to validity sections?"*.

Ideally, TTV sections should be a space where researchers communicate to other researchers reflections regarding the validity of their study. Such reflections should not be strictly bounded to a specific TTV categorization and its TTV types. While standards and checklists are not misguiding *per se*, following them mindlessly defies the original intent of TTV discussions. Without the conscious process of TTV knowledge creation, syntactically applying a TTV categorization does not, in most of the cases, accommodate the vast variety of research methods and TTV types characterizing a specific study.

Such reflection, however, should be done by reviewers of peer-reviewed papers as much as by their authors. Frequently, reviewers might ask authors to use such standards and checklists without questioning themselves their relevance and for what purpose. Consequently, our criticism is a call for action for both authors and reviewers.

Threats to validity have to be distinguished from the language used to reason about them **(P_6).** TTV, their categorizations, and types, are a common language used between researchers to communicate observations about what could have influenced the validity of a study. We should not be restrained by the TTV language itself, but rather acknowledge that we are using a language to express such concepts. Therefore, when outlining and reviewing TTV of a study, we should keep an open and creative mind, acknowledging that nuances, variations, or new threats that are not explicitly expressed in the current TTV language could have played a key role in a study.

Threats to validity need to be distinguished from limitations **(P_7).** We should recognize that TTV and limitations of a study are two distinct concepts. Researchers inspired by computer science design-science research tend to favor the term *limitations* while researchers inspired by empirical explanatory science favor the term *TTV*.

TTV are more precise, focusing on assessing the empirical strategies applied in the study to arrive at certain conclusions. Limitations is a broader concept, describing the scope of a study, focusing more on assessing the available options. One could say that TTV are the consequences of the choices made due to the limitations. In SE research, frequently design artifacts (such as tools, methods or interventions) are reported together with empirical investigations of the artifacts and/or the problem it claims to solve [2]. For

Table 1
Exemplifying the problems, and our vision on how to solve them

Problem	Current trend	Our vision
Threats to validity as checklists (P_1)	TTV sections based on "laundry list" guidelines tend to be shallow, lacking in depth analysis and discussion of threats in relation to the specific context and goal of the study at hand.	Checklist are used as a framework for insightful discussions of the research validity, helping to assess the value of the contribution.
Single threats to validity categorization (P_2)	Failing to acknowledge different types of TTV categorizations leads to blindly following the checklists that the authors are more accustomed, and not necessarily the best fitted for the study.	TTV analysis approaches are adapted to the main contributions of the study and the validity aspects of each of them that are most important to consider.
Widespread misunderstanding (P_3)	The research community, as both authors and reviewers, enforces an apparently systematic yet shallow TTV discussion. TTV guidelines and categorizations have to be strictly followed, even at the cost of discussing irrelevant threats. Relevant TTV which fall outside the considered TTV categorization are often disregarded.	TTV guidelines and categorizations are used as references, but do not strictly bind TTV discussions. TTV are freely analyzed by considering which are the most relevant threats for the study at hand, without a categorical restriction to a specific TTV framework.
Threats to validity as afterthought (P_4)	TTV is often considered in the final stage of a study, *i.e.,* documentation, and becomes a summary to justify the study in relation to the selected threats.	Integrate TTV as part of the study and consider it from the beginning. Especially, we need to explicitly evaluate and identify the threats in every phase of the study, mitigate and document them in an appropriate way.
Purpose of threats to validity (P_5)	TTV sections reported in many papers are too often limited to the description of the TTV types as expected by the readers (or reviewers). Sometimes they also miss the discussion of the related mitigations.	The true intent of a TTV section should be to help the reader (*e.g.,* the fellow researcher) deeply understand the extent to which the reported study can be applied to their own research, under which conditions, *e.g.,* to replicate it or extend it. This means that both TTV types and the discussion of applied mitigations may vary and may entail various levels of detail. Accordingly, we envision TTV sections to discuss both different purposes (*e.g.,* reuse, replication, extension) and TTV types and related mitigations which vary depending on the type of study and research method.
Threats to validity language (P_6)	TTV discussions are bounded to the language set by existing TTV guidelines and categorizations. Concepts which fall outside such systematic frameworks are disregarded, or considered as unreliable / irrelevant.	The TTV language is a dynamically adapted and evolving language used as a tool to communicate TTV section among researchers. The ultimate goal of such language is to reason about TTV concepts, without major concerns regarding a standardized syntactic adherence to the TTV language itself.
Threats to validity vs. limitations (P_7)	The uses of the terms TTV and limitations are not consistent in SE research. There is an overlap between the concepts and this overlap is treated differently by different groups of researchers.	A consistent distinction between TTV and limitations helping to gain a holistic perspective on the value of a research contribution.
Trade-offs between threats are not considered (P_8)	Discussing each type of TTV equally, trying to minimize them all at once, results in disconnection between the goal of the study and the validity of the outcome.	Different categories of TTV are traded against each other, to design the best study given its goals.

such studies it is important to distinguish between the limitations of the artifact as such and the TTV of the empirical investigations.

***Trade-offs among threats to validity are typically neglected* (P_8).** In the standardized use of TTV checklists, there is little room for discussions about trade-offs between different TTV types. Depending on the purpose of the research, some TTV can be accepted if others are mitigated. Siegmund *et al.* [6] observed lack of awareness if this, when surveying SE PC members.

Storey *et al.* [8] present a research framework, demonstrating that different empirical strategies have different potential research quality criteria, such as control, realism, and generalizability, which have to be traded. For example, a controlled classroom experiment reduces internal TTV, as the treatment and subjects are highly controlled. However, the external validity is limited, since real-world settings involve a lot more variation. Thus, the empirical strategy should be selected based on an upfront analysis of TTV in relation to study goals.

3. Concluding Vision

This section provides a glimpse on how problems P_1–P_8 identified in Section 2 could be addressed in future research. We use Table 1 to illustrate our vision, by mapping each current shortcoming of TTV sections (see column 'Current trend') to how it could be addressed in our vision (see column 'Our vision').

As an example, consider P_4 (Threats to validity as afterthought). The current trend is to analyze and discuss TTV retrospectively, i.e., authors often reflect after study execution on what they did or used in the study, their consequences, and how they impacted, or even determined, the study validity. In our vision, we could address P_4 by considering TTV in every phase of the study, including identifying potential threats, designing mitigation strategies, and documenting TTV appropriately.

We recognize there might be many ways to address the problems we identified. In this paper, we aim to start raising awareness, and critical reflection.

In future work, to analyze TTV-related problems, and study the potential impact recent SIGSOFT ESE standards

may have on TTVs, we plan to (i) perform a focused review on an exemplary literature set (*e.g.,* ICSE best papers), and (ii) complement results *via* focus groups with targeted communities (*e.g.,* the International Software Engineering Research Network). As end goal of a gradual research project, we strive to identify actions and solutions on how the current TTV issues can be concretely addressed and resolved.

References

[1] Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., Chatzigeorgiou, A., 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Information and Software Technology 106, 201–230. doi:10.1016/j.infsof.2018.10.006.

[2] Engström, E., Storey, M.A., Runeson, P., Höst, M., Baldassarre, M.T., 2020. How software engineering research aligns with design science: a review. Empirical Software Engineering 25, 2630–2660. doi:10.1007/s10664-020-09818-7.

[3] Feldt, R., Magazinius, A., 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. pp. 374–379.

[4] Maxwell, J., 1992. Understanding and Validity in Qualitative Research. Harvard Educational Review 62, 279–301. doi:10.17763/haer.62.3.8323320856251826.

[5] Petersen, K., Gencel, C., 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. pp. 81–89. doi:10.1109/IWSM-Mensura.2013.22.

[6] Siegmund, J., Siegmund, N., Apel, S., 2015. Views on internal and external validity in empirical software engineering, in: Bertolino, A., Canfora, G., Elbaum, S.G. (Eds.), 37th IEEE/ACM International Conference on Software Engineering, ICSE, Volume 1, IEEE Computer Society. pp. 9–19. doi:10.1109/ICSE.2015.24.

[7] Sjøberg, D.I.K., Bergersen, G.R., 2023. Construct validity in software engineering. IEEE Transactions on Software Engineering 49, 1374–1396. doi:10.1109/tse.2022.3176725.

[8] Storey, M.A., Ernst, N.A., Williams, C., Kalliamvakou, E., 2020. The who, what, how of software engineering research: a socio-technical framework. Empirical Software Engineering 25, 4097–4129. doi:10.1007/s10664-020-09858-z.

[9] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2012. Experimentation in software engineering. Springer Science & Business Media. doi:10.1007/978-3-642-29044-2.