Communicating Study Design Trade-offs in Software Engineering

MARTIN P. ROBILLARD, McGill University, Canada
DEEKSHA M. ARYA, McGill University, Canada
NEIL A. ERNST, University of Victoria, Canada
JIN L.C. GUO, McGill University, Canada
MAXIME LAMOTHE, Polytechnique Montréal, Canada
MATHIEU NASSIF, McGill University, Canada
NICOLE NOVIELLI, University of Bari, Italy
ALEXANDER SEREBRENIK, Eindhoven University of Technology, The Netherlands
IGOR STEINMACHER, Northern Arizona University, USA
KLAAS-JAN STOL, University College Cork and Lero, Ireland

Reflecting on the limitations of a study is a crucial part of the research process. In software engineering studies, this reflection is typically conveyed through discussions of study limitations or threats to validity. In current practice, such discussions seldom provide sufficient insight to understand the rationale for decisions taken before and during the study, and their implications. We revisit the practice of discussing study limitations and threats to validity and identify its weaknesses. We propose to refocus this practice of self-reflection to a discussion centered on the notion of *trade-offs*. We argue that documenting trade-offs allows researchers to clarify how the benefits of their study design decisions outweigh the costs of possible alternatives. We present guidelines for reporting trade-offs in a way that promotes a fair and dispassionate assessment of researchers' work.

CCS Concepts: • **Software and its engineering**;

ACM Reference Format:
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. 2023. Communicating Study Design Trade-offs in Software Engineering. *ACM Trans. Softw. Eng. Methodol.* 1, 1, Article 1 (January 2023), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

The software engineering research community employs a considerable diversity of research methods to advance the field [34, 35]. For instance, Stol and Fitzgerald cataloged 31 distinct terms used to refer to research methods in software engineering [34]. This diversity imposes a challenge on

Authors' addresses: Martin P. Robillard, robillard@acm.org, McGill University, Montréal, QC, Canada; Deeksha M. Arya, deeksha.arya@mail.mcgill.ca, McGill University, Montréal, QC, Canada; Neil A. Ernst, nernst@uvic.ca, University of Victoria, Victoria, BC, Canada; Jin L.C. Guo, jguo@cs.mcgill.ca, McGill University, Montréal, QC, Canada; Maxime Lamothe, maxime.lamothe@polymtl.ca, Polytechnique Montréal, Montréal, QC, Canada; Mathieu Nassif, mnassif@cs.mcgill.ca, McGill University, Montréal, QC, Canada; Nicole Novielli, nicole.novielli@uniba.it, University of Bari, Bari, Italy; Alexander Serebrenik, a.serebrenik@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands; Igor Steinmacher, igor.steinmacher@nau.edu, Northern Arizona University, Flagstaff, AZ, USA; Klaas-Jan Stol, k.stol@ucc.ie, University College Cork and Lero, Cork, Ireland.

ACM 1049-331X/2023/1-ART1
https://doi.org/XXXXXXX.XXXXXXX

researchers and reviewers to understand the implications of the use of various research methods in different contexts. Moreover, the design of a research study goes beyond selecting appropriate research methods and includes related research techniques and parameters [24]. The entire study design process is loaded with interconnected choices and dilemmas and constrained by available resources and other factors [19]. Each choice imposes *trade-offs* that impact the outcomes of a study and the interpretation of their results [21].

The notion of a trade-off in research study design is not new and ongoing discussion on the topic is taking place in many disciplines, including political and social science [4, 17, 20, 40]. Table 1 presents a selection of articles from a range of other disciplines that discuss the notion of threats to validity and trade-offs. Each discipline has its own specific considerations; for example, in the late 1980s, field research was not common within the Accounting discipline, and thus issues of validity and reliability of participant observation data became more prominent. Within Medicine, there is a stronger focus on alternative experimental designs, and strategies such as field studies do not loom large there. Despite these discipline-specific concerns, there is a common theme among these discussions, namely the need to consider alternative study designs and to make well-justified trade-offs.

While this topic is occasionally touched on in software engineering venues (e.g., [7, 32, 34, 39]), it is not currently common practice to systematically report trade-offs and their underlying reasoning when disseminating the research. We are also missing a shared practice for *how* to report trade-offs in study design. Such a gap limits our ability to assess the quality of research and can limit insights gained from replicating it.

FROM THREATS TO TRADE-OFFS

An extensive reflection on the decisions and dilemmas of a study's design is rarely reported in software engineering research. In sections dedicated to *threats to validity* or *limitations*, authors typically list the limitations of their study, following a rubric that includes construct, internal and external validity, and reliability, or its equivalent for qualitative studies [16]. Different studies have been dedicated to mapping threats to validity in specific subareas of software engineering, e.g., in secondary studies [1, 28, 41] or in software security research [5]. Despite these efforts to improve the reporting of threats to validity in software engineering research, a number of problems persist or even worsen as the practice intensifies [37]. We note three ISSUES in particular.

First, *threats to validity sections are often BOILERPLATE TEXT produced by rote.* For example, it is not unusual to find papers reporting a case study to declare that the study is limited to its context, and that the results cannot be generalized to other settings. Another common example is the limitation of sample size. Merely stating that sample size limits external validity is a truism. While it is usually a reminder that, indeed, the findings of a study must be considered in light of its limitations, these boilerplate statements simply highlight well-known essential limitations of given research methods [34] without providing additional insights.

Second, *existing focus on threats and limitations encourages a DEFENSIVE WRITING STYLE*, where the arguments often downplay the implications of a design choice out of fear of criticism [14]. Unfortunately, such an approach can obscure readers' view of the available design space for a decision. A particularly unfortunate instance of the defensive style is to rationalize study designs by referring to a previously published paper that also had made the same design decisions. This fragile reasoning relies on two major assumptions: first, that the authors of the cited study made a decision worth accepting without question, and second, that the decision remains valid outside its original context.

A third problem with threats to validity sections is that *they are OPAQUE ABOUT THE RATIONALE for some decisions that lead to study limitations.* For example, it can be unclear whether the limitations

Table 1. Selected discussions on threats to validity in other disciplines

Authors	Discipline	Description
Brutus and Duniewicz 2012 [3]	Political, Social and Behavioral Science	Reviews papers published in *The Leadership Quarterly* in 1990-2007, focusing on self-reported limitations. Main concerns include external validity and internal validity. Also discusses the dilemma to either disclose or not disclose limitations.
Capano and Engeli 2022 [4]	Policy Making	Discusses five methodological trade-offs including: parsimony, reliability, analytical purpose, comparative perspective, and performance assessment.
Lundberg 1997 [17]	Hospitality	Identifies the need to embrace a wider range of methods and to consider study design alternatives and trade-offs along each of the stages of a research study.
McKinnon 1988 [20]	Accounting	Presents considerations of threat to validity and reliability in field study research, which was novel at the time in accounting. Amongst others, presents trade-offs on different types of participant observations in terms of observer-caused effects, observer bias, data access, and complexities and limitations of the human mind.
McLaughlin and Talbert 1994 [21]	Education Research	Discusses a range of study design decisions and trade-offs, including scope, sample (e.g. many sites vs. few, purposive vs. random, embedded vs. distributed), and analytical perspective (longitudinal vs. cross-sectional, integrated vs. compartmentalized methods).
McNichols 2000 [22]	Earnings Management	Discusses trade-offs associated with research designs related to how company earnings are calculated. Presents evidence of possible misspecified earnings models, which can lead to wrong results, and proposals for how studies should report earning accruals.
Mercer et al. 2007 [23]	Preventive Medicine	Discusses research designs such as randomized controlled trial variants, quasi-experimental designs, and natural experiments, and trade-offs between them.
Smith et al. 1989 [33]	Entrepreneurship Research	Describes trade-offs around the type of data used (objective vs. subjective), and presents guidelines to use in selecting research methods.
Wanous et al. 1989 [38]	Applied Psychology	Highlights eight judgment calls and that based on these different researchers can come to different conclusions depending on the choices they make.
Wolff and Haase 2020 [40]	Comparative Urban Studies	Focuses on three trade-offs: analysis (reality vs. comparison); synthesis (comparison vs. theory), and description (reality vs. theory). Underpinning these trade-offs are the two contrasting poles that comparative urban studies focus on: the universal (generalizable) vs. the specific (concrete).

are the results of acceptable trade-offs. Even when limitations and their mitigation tactics are explained candidly and with technical details, they provide limited opportunity to understand to what extent they are justified.

We propose to evolve the existing practice of communicating threats to validity into a richly detailed commentary of the design trade-offs made for a study. This perspective introduces at least two additional aspects to study critiques commonly found in our field. First, a given study design decision can be assumed to be the most desirable of several possible *alternatives*. Second, a study design decision has *implications* that can be analyzed with respect to the alternatives. This

simple conceptual structure affords a multidimensional analysis of the impact of the study design decisions based on evidence. For example, a sensitivity analysis can provide quantitative support for a certain threshold used in a study. Qualitative evidence could include the results of a pilot study to support other study design decisions. Providing a detailed technical analysis of the trade-offs involved in a study design has several advantages: it helps justify a decision, both to the researchers themselves and to their readers; it provides reliable evidence and insights to assist researchers in designing future studies, and it promotes a rational and dispassionate approach to peer review by dissipating any expectation that some research design decisions can be superior in an absolute sense. Within this new format, the enumeration of threats to validity remains, but takes a new shape as *implications* of *study design decisions*.

COMMUNICATING TRADE-OFFS

Communicating a trade-off using a recognizable structure can help readers understand the underlying rationale, identify its important aspects, and use the results appropriately in future studies [29]. At the same time, research methods in software engineering are varied and so it is unlikely that a universal template will suit all purposes. We offer the following items as a structure to organize the presentation of a study design trade-off. The structure is a recommendation that need not be adopted strictly: rich insights are more valuable than indiscriminate conformance to a template.

We illustrate each item with corresponding fragments for a hypothetical trade-off description for the study *Turnover-Induced Knowledge Loss in Practice* [27]. This study *"sought to better understand the different contexts in which developers experience knowledge loss and the resulting implications."* The study relied on qualitative interviews with 27 professional developers and managers from three different software technology companies.

For convenience, we juxtapose the sample text fragments with the corresponding guidelines. We keep the example artificially short for presentation purposes. In practice, the items describing a trade-off should appear joined in a single paragraph and include more details. We provide two complete examples of trade-off descriptions in the next sections. In this and all other examples below, we use "we" to reflect the voice of the authors of the study cited. All examples are based on a study designed and conducted by at least one author of this correspondence.

A trade-off is identified by its **decision point**, which can act as its identifier. The description should include the value selected and be meaningful within a minimum of context, such as captured by the abstract of the paper.

A decision point is the number of companies to involve in the research. We involved three companies.

With a decision point come **alternatives**. It is key to consider the relative importance of these alternatives and how to organize them [24]. A review of alternatives can include properties, such as whether or not the set of alternatives is closed or whether or not they are mutually exclusive.

The alternatives were to focus on a single company, involve a small number (two or three), or sample many companies.

The selection of one alternative over competing options is the outcome of a system of **considerations** that relates the costs and benefits of each alternative, as well as constraints limiting the design space.

Increasing the number of of industry contexts comes at a cost associated with involving a company independently of the number of participants recruited from this company. This cost includes the effort to negotiate research agreements and to collect a sampling frame specific for the company.

The **rationale** for selecting an alternative can then be expressed in terms of these considerations. In some cases, it may be possible to distill a trade-off to a unidimensional cost-benefit equation that can even be quantified, while other trade-offs may be more complex. In any case, providing concrete evidence to support the rationale will help evaluate the study and to inform future work.

The rationale for involving three companies was two-fold. First, we wanted to study knowledge loss from at least two different company contexts, to support triangulation. Second, we wanted to involve a sufficient number of participants to meet common expectations for thematic analysis [11].

Additional discussion of the **implications** of the decision supports an in-depth exploration of the consequences of the choice made, in contrast to the inevitably more general cost-benefit calculus involved in the previous point (i.e., *considerations*). A separate item to address the implications can convey an assessment of the impact of the decision on the study's findings, including any threats to validity. Ideally, this assessment can include specific evidence (e.g., a sensitivity analysis).

Two of our four resulting themes are discussed by participants across all companies and the two other themes by participants from two companies. Thus, none of the themes is company-specific. Given our recruitment strategy, we expect that adding an additional company could have yielded around five additional participants (Companies B and C contributed four and six participants each, respectively).

The description of a trade-off can be enriched by any of the usual editorial devices, including references to literature and cross-references to relevant sections of the research report, and in particular to other trade-off descriptions. Threats to validity, now part of the implication section, can still be expressed using a familiar typology to ease the transition [31]. Alternatively, this transition can also be an opportunity to reassess and address the limitations of popular typologies [25, 37].

We expect that the trade-offs that most impact the research questions and findings will remain organized in a separate section. Within such a section, each trade-off can be described in its own titled paragraph or subsection, depending on its complexity. These trade-offs should relate to a significant aspect of the study design. Trade-offs of secondary importance, e.g., about a technical detail of the study environment, may be best located in proximity to the relevant context.

Including a detailed discussion of trade-offs inevitably requires additional space in a manuscript. As the examples below suggest, this space could amount to up to one page per major decision point. In situations where articles are limited in length or incur page charges, the relative value of a trade-off section will inevitably come into question. In such cases, the inclusion of a trade-off discussion itself becomes a trade-off. However, it can be argued that experimental trade-offs are a critical consideration in research methodology, and should therefore be represented in the article. When space is an issue, authors can be more concise than in the examples below while still addressing all of the main aspects of a trade-off. Further details can be relegated to an on-line appendix for completeness.

The trade-off structure we propose naturally addresses the three common limitations of *threats to validity* discussions. Because few experimental contexts are identical, calling for a context-based discussion of alternatives reduces the risk that common statements can be reused as BOILERPLATE TEXT across a majority of articles. Putting the emphasis on trade-offs also lifts the curtain on the design space that a DEFENSIVE WRITING STYLE can obscure. It is no longer a matter of arguing that the methodology chosen was the best one, but that the choice was reasonable and informed among a number of alternatives, each with their pros and cons. By the same token, our proposed structure specifically includes a *rationale* component to avoid being OPAQUE ABOUT THE RATIONALE for decisions that lead to limitations in a research design.

We now present two additional examples to illustrate our proposal in more detail.

EXAMPLE TRADE-OFF: RECRUITMENT APPROACH

For a study examining how developers respond to bots on GitHub [12], we required a sample of participants that represents the population of software developers who use pull requests.

The **decision point** stems from determining from where the population of participants will be recruited. Recruiting software developers is challenging as they are a specialist population who can be hard to contact [10, 36]. Many factors can come into play when deciding on a sampling frame for this population [2, 10].

The **alternatives** were 1) to recruit developers from Prolific,[1] an online platform that offers a pool of study participants and tools for managing payment and other study operations; 2) to recruit students via university channels; 3) to approach developers directly using GitHub profiles; 4) to choose another crowd-worker platform such as Mechanical Turk (MTurk);[2] or 5) a combination of these. Our decision was to recruit developers from Prolific, and also from students in our classes.

The **rationale** for selecting a combination of students and Prolific can be explained in light of the following **considerations**:

(1) *Prolific participants are paid.* This likely incentivizes participation from people outside the required target population. *Prolific participants are easier to recruit*, as Prolific handles recruitment, screening, and compensation directly. The platform allows for a set of filters ensuring, for example, that an equal number of women and men are selected. *Prolific participants also need to be screened* with a challenging set of filtering questions to ensure competence on the task [6]. We did not evaluate MTurk in detail; previous software engineering studies and our own experience with Prolific had been mostly positive [8, 26, 30], and MTurk participants are usually involved in different activities, such as data annotations, and thus not dedicated to participating in user studies.
(2) As for considering only students, this would have *limited the insights* to what upper-year undergraduates perceive with bots, and made *recruitment more challenging* as few students have experience with GitHub bots.
(3) Direct recruitment of developers is costly due to the overhead of identifying and contacting potential recruits and of the low expected yield [9] although with good planning this can be mitigated [10]. The sample is also self-selected to people with interest in the study. On the positive side, validity may be higher, as participants would clearly be members of the target population.

One **implication** of our choice was that we lost a large number of initial volunteers who failed the Prolific screening. Only 12% were able to complete the study. However, rigorous screening (and attention questions) gave us more confidence in the validity of the results.

Another implication of using crowd-worker platforms is study costs. Even ineligible volunteers still require payment for their time. For every 100 volunteer for our study on Prolific we paid approximately 42 USD plus service fees for unusable data. These fees could be seen as a price to pay to ensure validity of the data. Furthermore, if the survey instrument is flawed (for example, it does not correctly randomize for order effects), participants who complete the study are still paid for this unusable data. Before beginning the study, there is an unknown total cost per usable data point collected, even with an upper bound on total costs [36].

[1] https://www.prolific.co/. Verified 2024-02-22.
[2] https://www.mturk.com/. Verified 2024-02-22.

EXAMPLE TRADE-OFF: DATA COLLECTION APPROACH

> *For a study investigating developers' emotions and perceived productivity [13], we needed a reliable and effective way to measure developers' productivity. Developer productivity is a complex phenomenon that cannot be measured by a metric in isolation. Productivity depends not only on the activity that one is working on but also on personal, task, and team contexts.*

One **decision point** concerned the mechanism used to collect data on developer productivity. Our decision was to use self-reported productivity. Since the goal of this study requires a relationship of trust with the participants, we considered it appropriate to ask them to self-report their perceived productivity for each activity they were working on during the study. An **alternative** to self-reporting would have been to use telemetry, for example by instrumenting the participants' computers with activity trackers, or by analyzing metrics associated with their development activities, such as their number of commits.

We **considered** the level of invasiveness to participants required to collect data [15]. The trade-off is privacy vs. more details about productivity. An approach of implicit data collection via telemetry would provide us with more details about the tasks performed by each developer to collect data to enable measuring their performance. However, this approach invades developers' privacy and could influence their behavior. By adopting a self-reporting mechanism, it is possible to ask the developers to self-assess and report their productivity using a sampling approach. Using the self-reporting approach allows participants to report on their perception of productivity without the researchers having access to other information that may be considered sensitive to them. In this case, they may report inaccurate productivity because of, for example, fear of judgment. Still, self-reporting involves interrupting developers during their work and this may trigger negative emotions and affect the results. Revealing productivity levels and emotions might also be difficult for developers to disclose in a professional setting [18]. Ultimately, trust and other personal aspects of this research were especially important, ruling out the use of invasive telemetry instrumentation, which was the main **rationale** for our decision.

The **implication** of collecting self-reported data is that the number of observations was limited to one report every 60 minutes, thus resulting in far fewer data points than using telemetry. Fortunately, we noted that the participants were eager to contribute to the research by self-reporting their emotional states, their causes for them, and their self-perceived productivity. Even though provided with the possibility to dismiss the pop-up questionnaire, on average, participants filled the questionnaire 5.4 times per day. The days for which the self-reports were missing were mainly due to participants not being at work for personal reasons. Overall, the participants did not self-report without explanation only for five days (out of 42).

CONCLUSION

Designing a research study unavoidably involves trade-offs that can impact its results. Concentrating on consequent limitations or threats to validity, instead of researchers' deliberations before and during a study, can result in defensive boilerplate that provides little insight into the reasoning for design decisions. We propose an open discussion of study design trade-offs that includes alternatives, considerations made, and detailed implications on the outcome. We demonstrate the application of our proposal with three examples from our prior research studies. It is our hope that by encouraging the inclusion of these trade-offs in calls for papers, by including them in our own future research papers, and by carefully assessing them when we act as reviewers, our community can enable more faithful replication of research and further encourage open science.

ACKNOWLEDGMENTS

This article is an outcome of the Symposium on Empirical Software Engineering held at the Bellairs Research Institute in January 2023. The authors are grateful to Gema Rodríguez-Pérez, Bogdan Vasilescu, and Shurui Zhou for discussions, and to the reviewers for additional feedback. This work is supported by the Natural Sciences and Engineering Research Council of Canada, the European Union—NextGenerationEU through the Italian Ministry of University (PRIN 2022), the US National Science Foundation, and the Science Foundation Ireland.

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] Sebastian Baltes and Paul Ralph. 2022. Sampling in software engineering research: a critical review and guidelines. *Empirical Software Engineering* 27, 4, Article 94 (2022), 31 pages. https://doi.org/10.1007/s10664-021-10072-8

[3] Stéphane Brutus and Kris Duniewicz. 2012. The many heels of Achilles: An analysis of self-reported limitations in leadership research. *The Leadership Quarterly* 23, 1 (2012), 202–212. https://doi.org/10.1016/j.leaqua.2011.11.015

[4] Giliberto Capano and Isabelle Engeli. 2022. Using Instrument Typologies in Comparative Research: Conceptual and Methodological Trade-Offs. *Journal of Comparative Policy Analysis: Research and Practice* 24, 2 (2022), 99–116. https://doi.org/10.1080/13876988.2020.1871297

[5] Daniela S. Cruzes and Lofti ben Othmane. 2017. Threats to Validity in Empirical Software Security Research. In *Empirical Research for Software Security: Foundations and Experience*, Lofti ben Othmane, Martin Gilje Jaatun, and Edgar Weippl (Eds.). CRC Press, Chapter 10, 275–300. https://doi.org/10.1201/9781315154855-10

[6] Anastasia Danilova, Alena Naiakshina, Stefan Horstmann, and Matthew Smith. 2021. Do you Really Code? Designing and Evaluating Screening Questions for Online Surveys with Programmers. In *Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering*. 537–548. https://doi.org/10.1109/icse43902.2021.00057

[7] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. Selecting Empirical Methods for Software Engineering Research. In *Guide to Advanced Empirical Software Engineering*, Forrest Shull, Janice Singer, and Dag I. K. Sjøberg (Eds.). Springer, Chapter 11, 285–311. https://doi.org/10.1007/978-1-84800-044-5_11

[8] Felipe Ebert, Alexander Serebrenik, Christoph Treude, Nicole Novielli, and Fernando Castor. 2022. On Recruiting Experienced GitHub Contributors for Interviews and Surveys on Prolific. (2022), 3 pages pages. Retrieved 2023-07-07 from https://www.win.tue.nl/~aserebre/ROPES2022.pdf

[9] Davide Falessi, Natalia Juristo, Claes Wohlin, Burak Turhan, Jürgen Münch, Andreas Jedlitschka, and Markku Oivo. 2018. Empirical software engineering experts on the use of students and professionals in experiments. *Empirical Software Engineering* 23, 1 (2018), 452–489. https://doi.org/10.1007/s10664-017-9523-3

[10] Robert Feldt, Thomas Zimmermann, Gunnar R. Bergersen, Davide Falessi, Andreas Jedlitschka, Natalia Juristo, Jürgen Münch, Markku Oivo, Per Runeson, Martin Shepperd, Dag I. K. Sjøberg, and Burak Turhan. 2018. Four commentaries on the use of students and professionals in empirical software engineering experiments. *Empirical Software Engineering* 23, 6 (2018), 3801–3820. https://doi.org/10.1007/s10664-018-9655-0

[11] Andrew J. B. Fugard and Henry W. W. Potts. 2015. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. *International Journal of Social Research Methodology* 18, 6 (2015), 669–684. https://doi.org/10.1080/13645579.2015.1005453

[12] Amir Ghorbani, Nathan Cassee, Derek Robinson, Adam Alami, Neil A. Ernst, Alexander Serebrenik, and Andrzej Wąsowski. 2023. Autonomy Is An Acquired Taste: Exploring Developer Preferences for GitHub Bots. In *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering*. 13 pages pages.

[13] Daniela Girardi, Filippo Lanubile, Nicole Novielli, and Alexander Serebrenik. 2022. Emotions and Perceived Productivity of Software Developers at the Workplace. *IEEE Transactions on Software Engineering* 48, 9 (2022), 3326–3341. https://doi.org/10.1109/TSE.2021.3087906

[14] Ethan M. Higgins. 2018. The State of Peer Review in Criminology: Literary Theory, Perceptions, and the Catch-22 Metaphor of Peer Review. *Journal of Criminal Justice Education* 29, 4 (2018), 507–530. https://doi.org/10.1080/10511253.2017.1420809

[15] Timothy C. Lethbridge, Susan Elliott Sim, and Janice Singer. 2005. Studying Software Engineers: Data Collection Techniques for Software Field Studies. *Empirical Software Engineering* 10, 3 (2005), 311–341. https://doi.org/10.1007/s10664-005-1290-x

[16] Yvonna S. Lincoln and Egon G. Guba. 1985. *Naturalistic Inquiry*. Sage Publications.

[17] Craig C. Lundberg. 1997. Widening the Conduct of Hospitality Inquiry: Toward Appreciating Research Alternatives. *Journal of Hospitality & Tourism Research* 21, 1 (1997), 1–13. https://doi.org/10.1177/109634809702100103

[18] David Matsumoto. 1990. Cultural Similarities and Differences in Display Rules. *Motivation and Emotion* 14, 3 (1990), 195–214. https://doi.org/10.1007/BF00995569

[19] Joseph E. McGrath. 1981. Dilemmatics: The Study of Research Choices and Dilemmas. *American Behavioral Scientist* 25, 2 (1981), 179–210. https://doi.org/10.1177/000276428102500205

[20] Jill McKinnon. 1988. Reliability and Validity in Field Research: Some Strategies and Tactics. *Accounting, Auditing & Accountability Journal* 1, 1 (1988), 34–54. https://doi.org/10.1108/EUM0000000004619

[21] Milbrey W. McLaughlin and Joan E. Talbert. 1994. School Context Research: Design Choices, Tradeoffs and Payoffs. *The Australian Educational Researcher* 21, 2 (1994), 63–85. https://doi.org/10.1007/BF03219568

[22] Maureen F. McNichols. 2000. Research design issues in earnings management studies. *Journal of Accounting and Public Policy* 19, 4–5 (2000), 313–345. https://doi.org/10.1016/S0278-4254(00)00018-1

[23] Shawna L. Mercer, Barbara J. DeVinney, Lawrence J. Fine, Lawrence W. Green, and Denise Dougherty. 2007. Study Designs for Effectiveness and Translation Research: Identifying Trade-offs. *American Journal of Preventive Medicine* 33, 2 (2007), 139–154. https://doi.org/10.1016/j.amepre.2007.04.005

[24] Theodore J. Mock. 1972. A Decision Tree Approach to the Methodological Decision Process. *The Accounting Review* 47, 4 (1972), 826–829.

[25] Charles S. Reichardt. 2011. Criticisms of and an alternative to the Shadish, Cook, and Campbell validity typology. *New Directions for Evaluation* 2011, 130 (2011), 43–53. https://doi.org/10.1002/ev.364

[26] Brittany Reid, Markus Wagner, Marcelo d'Amorim, and Christoph Treude. 2022. Software Engineering User Study Recruitment on Prolific: An Experience Report. In *1st International Workshop on Recruiting Participants for Empirical Software Engineering (RoPES'22)*. 3 pages pages. arXiv:2201.05348

[27] Martin P. Robillard. 2021. Turnover-Induced Knowledge Loss in Practice. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1292–1302. https://doi.org/10.1145/3468264.3473923

[28] Martin P. Robillard, Mathieu Nassif, and Shane McIntosh. 2018. Threats of Aggregating Software Repository Data. In *Proceedings of the 34th IEEE International Conference on Software Maintenance and Evolution*. 508–518. https://doi.org/10.1109/ICSME.2018.00009

[29] Paula T. Ross and Nikki L. Bibler Zaidi. 2019. Limited by our limitations. *Perspectives on Medical Education* 8, 4 (2019), 261–264. https://doi.org/10.1007/s40037-019-00530-x

[30] Daniel Russo. 2022. Recruiting Software Engineers on Prolific. In *1st International Workshop on Recruiting Participants for Empirical Software Engineering (RoPES'22)*. 2 pages pages. arXiv:2203.14695

[31] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

[32] Janet Siegmund, Norbert Siegmund, and Sven Apel. 2015. Views on internal and external validity in empirical software engineering. In *Proceedings of the 37th IEEE/ACM International Conference on Software Engineering*, Vol. 1. 9–19. https://doi.org/10.1109/ICSE.2015.24

[33] Ken G. Smith, Martin J. Gannon, and Harry J. Sapienza. 1989. Selecting Methodologies for Entrepreneurial Research: Trade-offs and Guidelines. *Entrepreneurship Theory and Practice* 14, 1 (1989), 39–50. https://doi.org/10.1177/104225878901400104

[34] Klaas-Jan Stol and Brian Fitzgerald. 2018. The ABC of Software Engineering Research. *ACM Transactions on Software Engineering and Methodology* 27, 3, Article 11 (2018), 51 pages. https://doi.org/10.1145/3241743

[35] 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

[36] Mohammad Tahaei and Kami Vaniea. 2022. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3491102.3501957

[37] Roberto Verdecchia, Emelie Engström, Patricia Lago, Per Runeson, and Qunying Song. 2023. Threats to validity in software engineering research: A critical reflection. *Inf. Softw. Technol.* 164 (2023), 107329. https://doi.org/10.1016/J.INFSOF.2023.107329

[38] John P. Wanous, Sherry E. Sullivan, and Joyce Malinak. 1989. The Role of Judgment Calls in Meta-Analysis. *Journal of Applied Psychology* 74, 2 (1989), 259–264. https://doi.org/10.1037/0021-9010.74.2.259

[39] Claes Wohlin and Aybüke Aurum. 2015. Towards a decision-making structure for selecting a research design in empirical software engineering. *Empirical Software Engineering* 20, 6 (2015), 1427–1455. https://doi.org/10.1007/s10664-014-9319-7

[40] Manuel Wolff and Annegret Haase. 2020. Viewpoint: Dealing with trade-offs in comparative urban studies. *Cities* 96, Article 102417 (2020), 7 pages. https://doi.org/10.1016/j.cities.2019.102417

[41] 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 *Proceedings of the 23rd Asia-Pacific Software Engineering Conference*. 153–160. https://doi.org/10.1109/APSEC.2016.031