ACM SIGSOFT
Empirical Standards

Version 0.1.0

Recommended citation: Paul Ralph et al. (2020) ACM SIGSOFT Empirical Standards. arXiv:2010.03525 [cs.SE].

OCTOBER 14

Paper and Peer Review Quality Task Force
Edited by: Paul Ralph

https://github.com/acmsigsoft/EmpiricalStandards

SIGSOFT

Table of Contents

Introduction

Why are so many manuscripts rejected by peer review? Constant rejection is neither intrinsic to science nor necessary for quality control. Rather, constant rejection is rooted in dissensus within scientific communities regarding how research should be conducted. The *ACM SIGSOFT Paper and Peer Review Quality Initiative* seeks to increase paper quality, review quality, acceptance rates and consensus around research practices by generating and evolving *empirical standards.*

> *Empirical Standard: A brief public document that communicates expectations for a specific kind of study (e.g. a questionnaire survey).*

Empirical standards **are not** vague criteria like "soundness" and "presentation." Empirical standards contain lists of specific practices or attributes related to a particular methodology, e.g., "uses random assignment" (experiment) or "presents clear chain of evidence from interviewee quotations to proposed concepts" (qualitative survey).

Empirical standards **do not** replace expert judgment with inflexible rubrics. For example, if researchers do not report effect sizes with confidence intervals because they took a Bayesian approach and report posterior probabilities instead, they just say so. Furthermore, empirical standards focus on the methodological substance of a study; they do not micromanage style.

Empirical standards allow crucial decisions about what is and is not acceptable scientific practice to be made by a community, collectively, rather than by a reviewer, individually. For example, we, as a community should decide whether a single in-depth case study is sufficient for a full-length technical article. This decision should not be made by individual reviewers, case-by-case.

The standards presented below were generated by a large, diverse team of domain experts (see **List of Contributors**). However, they are not final. The standards are meant to reflect the views of our community and to evolve as those views evolve. It is crucial that the standards are fair and inclusive. Any active software engineering researcher can suggest improvements and we are working on a governance model for updating the standards regularly.

The best way to understand what we mean by an empirical standard is to review the draft standards themselves (next), perhaps beginning with a standard for a familiar methodology. Note that we try to avoid duplication, so, for example, the *Case Study Standard* does not say "states a clear research question" because that is already in the *General Standard*. After the standards, we include several supplements for cross-cutting concerns such information visualization and sampling. We then explain how they might be used, discuss their benefits and relate how they were generated.

The General Standard

Application

This general standard applies to all software engineering studies that collect and analyze data. It should be complemented by more specific guidelines where available.

Initial Checks (Editor)

Reviewers should only be invited for papers with the following attributes. By assigning reviewers, the editor/chair/administrator is confirming that the manuscript meets these criteria:

- ☐ meets venue's requirements (e.g. length, author-blinding, appropriate keywords)
- ☐ within the venue's scope
- ☐ meets the minimum level of language quality acceptable to the journal
- ☐ cites other scholarly works
- ☐ presents new analysis not previously published in a peer-reviewed venue (i.e. preprints are fine)
- ☐ does not include unattributed verbatim published text (i.e. plagiarism)

Initial Checks (Reviewer)

Before beginning to review a paper, assigned reviewers should verify the following.

- ☐ reviewer has no conflicts of interest; if unsure, check with the chair or editor
- ☐ reviewer has sufficient expertise; if unsure, check with the chair or editor and clarify what you can(not) evaluate
- ☐ paper is clear enough (in language and presentation) to even review

Specific Attributes

Importance	Attribute
Essential	☐ states a purpose, problem, objective, or research question
	☐ methodology is appropriate (not necessarily optimal) for stated purpose or questions
	☐ describes in detail *what*, *where*, *when* and *how* data were collected
	☐ describes in detail how the data were analyzed
	☐ discusses and validates assumptions of any statistical tests used.
	☐ presents results; results directly address research questions
	☐ discusses the importance, implications and limitations (validity threats) of the study
	☐ contributes in some way to the collective body of knowledge (see **Replications Supplement**)
	☐ supports claims and conclusions with explicit arguments or evidence (data/observations)
	☐ defines jargon, acronyms and key concepts
	☐ language is not misleading; any grammatical problems do not substantially hinder understanding
	☐ visualizations/graphs are not misleading (see the **Information Visualization Supplement**)
	☐ complies with all applicable **empirical standards**
Desirable	☐ summarizes and synthesizes a reasonable selection of related work
	☐ clearly describes relationship between contribution(s) and related work
	☐ states epistemological stance (e.g. post-positivism, interpretivism, critical realism)
	☐ appropriate statistical power (for quantitative work) or saturation (for qualitative work)
	☐ reasonable attempts to investigate or mitigate limitations
	☐ discusses study's realism, assumptions and sensitivity of the results to the realism/assumptions
	☐ provides plausibly useful interpretations or recommendations for practice, education or research
	☐ openly shares data and materials to the extent possible within practical and ethical limits
	☐ concise, precise, well-organized and easy-to-read presentation
	☐ visualizations (e.g. graphs, diagrams, tables) advance the paper's arguments or contribution
	☐ clarifies the roles and responsibilities of the researchers (i.e. who did what?)
	☐ provides an auto-reflection or assessment of the authors' own work (e.g. lessons learned)
Extraordinary	☐ applies two or more data collection or analysis strategies to the same research question (see **Multimethodology Supplement**)
	☐ approaches the same research question(s) from multiple epistemological perspectives
	☐ innovates on research methodology while completing an empirical study

General Quality Criteria

There are no universal quality criteria. Each study should be assessed against quality criteria appropriate for its methodology, as laid out in the specific **empirical standards**. Avoid applying inappropriate quality criteria (e.g. construct validity to a study with no constructs; internal validity to a study with no causal relationships).

Examples of Acceptable Deviations

A study can only apply an **empirical standard** if an appropriate standard exists. If no related standards exist, studies should apply published guidance. If no appropriate guidance exists, reviewers should apply the general standard and construct an ad hoc evaluation scheme for the new method.

Good Review Practices

Reviewers evaluate a manuscripts' trustworthiness, importance and clarity. The results must be, primarily, true (trustworthy) and, secondarily, important. A paper that is trustworthy can be accepted even if it is not important. A paper that is not trustworthy cannot be accepted, even if it seems important. Papers that are both trustworthy and important can have priority. Papers must be clear enough to judge their trustworthiness and importance. Reviewers should endeavor to:

- Reflect on and clearly state their own limitations and biases.
- Clarify which are necessary and which are suggested changes. Ideally, separate them.
- Identify parts of the paper that you cannot effectively judge or did not review.

Invalid Criticisms and Reviewing Antipatterns

- Applying empirical standards in a mechanical, inflexible, box-ticking or gotcha-like manner.
- Rejecting a study because it uses a methodology for which no specific standard is available.
- Skimming a manuscript instead of carefully reading each word and inspecting each figure and table.
- Unprofessional or vitriolic tone, ad hominem attacks, disparaging or denigrating comments.
- Allowing the authors' identities or affiliations to affect the review.
- Focusing on superficial details of paper without engaging with its main claims or results.
- Stating that a study: (i) lacks detail without enumerating missing details; (ii) is of low quality without explaining specific problems; or (iii) is not new without providing citations to published studies that make the same contribution.
- Criticizing a study for limitations intrinsic to that kind of study or the methodology used.
- Cross paradigmatic criticism (e.g. attacking an interpretivist study for not conforming to positivist norms).
- Using sub-reviewers when the venue does not explicitly allow it.
- Using the review to promote the reviewer's own views, theories, methods, or publications.
- Rejecting a study because the reviewer would have used a different methodology or design.

Research and Reporting Antipatterns

- Attempting a study without reading, understanding and applying published guidelines for that kind of study.
- Unreasonably small, underpowered or limited studies.
- Hypothesizing After Results are Known (HARKING) in ostensibly confirmatory, (post-)positivist research.
- Reporting only the subset of statistical tests that produce significant results (p-hacking).
- Reporting—together in one paper—several immature or disjointed studies instead of one fully-developed study.
- Unnecessarily dividing the presentation of a single study into many papers (salami-slicing).
- Overreaching conclusions or generalizations; obfuscating, downplaying or dismissing a study's limitations.
- Mentioning related work only to dismiss it as irrelevant; listing rather than analyzing and synthesizing related work.

Engineering Methods

Engineering Research

Research that invents and evaluates technological artifacts

Application

This standard applies to manuscripts that propose and evaluate technological artifacts, including algorithms, models, languages, methods, systems, tools, and other computer-based technologies. This standard is not appropriate for:

- evaluations of pre-existing engineering research approaches (consider the Experiments Standard)
- experience reports of applying pre-existing engineering research approaches

Specific Attributes

Importance	Attribute
Essential	☐ describes the proposed artifact in adequate detail[1]
	☐ justifies the need for, usefulness of, or relevance of the proposed artifact[2]
	☐ conceptually evaluates the artifact; discusses its strengths, weaknesses and limitations[3]
	☐ EITHER: discusses state-of-art alternatives (and their strengths, weaknesses and limitations) OR: explains why no state-of-art alternatives exist OR: provides compelling argument that direct comparisons are impractical
	☐ Empirically evaluates the proposed artifact using: **action research**, in which the researchers intervene a real organization using the artifact, a **case study** in which a real organization uses the artifact without researcher intervention, a **controlled experiment** in which human participants use the artifact, a **simulation** in which the artifact is used in an artificial environment, or another method for which a clear and convincing rationale is provided
	☐ clearly indicates the empirical methodology being used (e.g. action research, controlled experiment)
	☐ EITHER: empirically compares the artifact to one or more state-of-the-art alternative artifacts OR: empirically compares the artifact to one or more state-of-the-art benchmarks OR: provides a clear and convincing rationale for why comparative evaluation is impractical
	☐ assumptions (if any) are explicit; do not contradict each other or the contribution's goals; plausibly hold for the evaluation subjects
	☐ uses notation consistently (if any notation is used)
Desirable	☐ reviews the theoretical basis of the artifact
	☐ provides correctness arguments of the key analytical and theoretical contributions (e.g. theorems, complexity analyses, mathematical proofs)
	☐ includes one or more running examples to elucidate the artifact
	☐ evaluates the artifact in an industry-relevant context (e.g. widely used open-source projects, professional programmers)
	☐ provides a replication package including datasets and analytical scripts and EITHER a comprehensive description of the artifact OR source code if artifact is virtual
	☐ justifies any items missing from replication package based on practical or ethical grounds.
Extraordinary	☐ contributes to our collective understanding of design practices or principles
	☐ presents ground-breaking innovations with obvious real-world benefits

General Quality Criteria

- Comprehensiveness of proposed artifact description
- Appropriateness of evaluation methods to the nature, goals, and assumptions of the contribution
- Relationship of innovativeness to rigorousness: less innovative artifacts require more rigorous evaluations

[1] e.g., does the paper describe the overall workflow of the solution, showing how different techniques work together? Are algorithmic contributions presented in an unambiguous way? Are the key parts of a formal model presented explicitly? Are the novel components of the solution clearly singled out?

[2] i.e., is the problem the proposed approach tries to solve specific to a certain domain? If so, why? Why are state-of-the-art approaches not good enough to deal with the problem? How can the technical contribution be beneficial?

[3] e.g., time complexity of an algorithm; theoretical

Antipatterns

- overstates the novelty of the contribution
- omits details of key conceptual aspects while focusing exclusively on incidental implementation aspects
- evaluation consists *only* of eliciting users' opinions of the artifact
- evaluation consists *only* of quantitative performance data that is not compared to established benchmarks or alternative solutions (see related point in "Invalid Criticism")

Invalid Criticisms

- The paper does not report as ambitious an empirical study as other predominately empirical papers. The more innovative the artifact and more comprehensive the conceptual evaluation, the less we should expect from the empirical study.
- Too few experimental subjects (e.g. the source code used to evaluate a static analysis technique) if few subjects are available in the contribution's domain or the experimental evaluation is part of a more comprehensive validation strategy (e.g. formal arguments). Other criteria, such as the variety, realism, availability, and scale of the subjects, should also be considered to assess the quality of the evaluation.
- No replication package, if there are clear, convincing practical or ethical reasons preventing artifact disclosure.
- The artifact is not experimentally compared with related approaches *that are not publicly available*. In other words, before saying "you should have compared this against X, make sure X is actually available and functional.
- This is not the first known solution to the identified problem. The novelty of the paper can be in how it achieves scalability, better performance on specific classes of problems, applicability to realistic systems, stronger theoretical guarantees, or other aspects of improvement. Proposed artifacts should outperform existing artifacts on *some* dimension(s).
- The contribution is not technically complicated. What matters is that it works. Unnecessary complexity is undesirable.

Suggested Readings[4]

Richard Baskerville, Jan Pries-Heje, and John Venable. 2009. Soft design science methodology. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (DESRIST '09).* Association for Computing Machinery, New York, NY, USA, Article 9, 1–11. DOI: 10.1145/1555619.1555631

Carlo Ghezzi. 2020. *Being a researcher - an informatics perspective*. Springer Nature.

Alan Hevner and Samir Chatterjee. 2010. *Design Research in Information Systems*. Integrated Series in Information Systems. Springer, 22, (Mar. 2010), 145-156. DOI: 10.1007/978-1-4419-5653-8_11

Alan R. Hevner, Salvatore T. March, Jinsoo Park and Sudha Ram. 2004. Design Science in Information Systems Research. *MIS Quarterly*, 28, 1 (Mar. 2004), 75–105. DOI:10.2307/25148625.

Roel Wieringa. 2014. *Design science methodology for information systems and software engineering.* Springer.

Exemplars

Kihong Heo, Hakjoo Oh and Hongseok Yang. 2019. Resource-aware Program Analysis via Online Abstraction Coarsening. In *Proceedings of the 41st International Conference on Software Engineering.*

Jianhui Chen, Fei He. 2018. Control Flow-Guided SMT Solving for Program Verification. In *Proceedings of the 33rd International Conference on Automated Software Engineering*.

Calvin Loncaric, Michael D. Ernst and Emina Torlak. 2018. Generalized Data Structure Synthesis. In *Proceedings of the 40th International Conference on Software Engineering.*

Nikolaos Tsantalis, Davood Mazinanian and Shahriar Rostami Dovom. 2017. Clone Refactoring with Lambda Expressions. In *Proceedings of the 39th International Conference on Software Engineering.*

August Shi, Suresh Thummalapenta, Shuvendu Lahiri, Nikolaj Bjorner and Jacek Czerwonka. (2017) Optimizing Test Placement for Module-Level Regression Testing. In *Proceedings of the 39th International Conference on Software Engineering.*

Magnus Madsen, Frank Tip, Esben Andreasen, Koushik Sen, and Anders Møller. 2016. Feedback-Directed Instrumentation for Deployed JavaScript Applications. In *Proceedings of the 38th International Conference on Software Engineering*

[4] Note: some of the following readings incorrectly refer to engineering research as "design science" because the information systems community uses this term. Design science properly refers to the study of designers and their processes. Learning by building innovative artifacts is more correctly called engineering research.

Qualitative Methods

Action Research

Empirical research that investigates how an intervention, like the introduction of a method or tool, affects a real-life context

Application

This standard applies to empirical research that meets the following conditions.

- investigates a primarily social phenomenon within its real-life, organizational context
- intervenes in the real-life context (otherwise see the **Case Study Standard**)
- the change and its observation are an integral part of addressing the research question and contribute to research

If the intervention primarily alters social phenomena (e.g. the organization's processes, culture, way of working or group dynamics), use this standard. If the intervention is a new technology or technique (e.g. a testing tool, a coding standard, a modeling grammar), especially if it lacks a social dimension, consider the **Engineering Research Standard**. If the research involves creating a technology and an organizational intervention with a social dimension, consider both standards.

Specific Attributes

Importance	Attribute
Essential	☐ describes the context or site of the intervention(s)
	☐ describes the intervention(s) in detail
	☐ describes the relationship between the researcher and the host organization[5]
	☐ reports the length of the project and describes the longitudinal dimension of the research design
	☐ describes the interactions between researcher(s) and host organization(s)—what the interventions were, who intervened and with which part of the organization, as well as the outcome of the interventions
	☐ describes how interventions were determined (by management, researchers, or participative/co-determination process)
	☐ explains research cycles or phases, if any, and their relationship to the intervention(s)[6]
	☐ explains how the interventions are evaluated[7]
	☐ reports participant or stakeholder reactions to interventions
	☐ presents a clear and well-argued chain-of-evidence from observations to findings
	☐ reports lessons learned by the organization
	☐ researchers reflect on their own possible biases
Desirable	☐ uses direct quotations extensively
	☐ uses member checking to assess resonance
	☐ findings plausibly transferable to other contexts
	☐ triangulation across quantitative and qualitative data
Extraordinary	☐ research team with triangulation across researchers (to mitigate researcher bias)

General Quality Criteria

Example criteria include reflexivity, credibility, resonance, usefulness and transferability (see **Glossary**). Positivist quality criteria such as internal validity, construct validity, generalizability and reliability typically do not apply.

Examples of Acceptable Deviations

- In a study of deviations from organizational standards, detailed description of circumstances and direct quotations are omitted to protect participants.
- The article reports a negative outcome of an intervention and e.g. investigates why a certain method was not applicable.

Antipatterns

- Forcing interventions that are not acceptable to participants or the host organization.
- Losing professional distance; becoming unable to evaluate the intervention impartially; going native.
- Over-selling a tool or method without regard for participants' problems, practices or values.

[5] E.g. project financing, potential conflicts of interest, professional relationship leading to access

[6] Action research projects are structured in interventions often described as action research cycles, which are often structured in distinct phases. It is a flexible methodology, where subsequent cycles are based on their predecessors.

[7] Can include quantitative evaluation in addition to qualitative evaluation.

- Avoiding systematic evaluation; downplaying problems; simply reporting participants views of the intervention.

Invalid Criticisms

- The findings and insights are not valid because the research intervened in the context. Though reflexivity is crucial, the whole point of action research is to introduce a change and observe how participants react.
- This is merely consultancy or an experience report. Systematic observation and reflection should not be dismissed as consultancy or experience reports. Inversely, consultancy or experiences should not be falsely presented as action research.
- Lack of quantitative data; causal analysis; objectivity, internal validity, reliability, or generalizability.
- Sample not representative; lack of generalizability; generalizing from one organization.
- Lack of replicability or reproducibility; not releasing transcripts.
- Lack of control group or experimental protocols. An action research study is not an experiment.

Suggested Readings

Richard Baskerville and A. Trevor Wood-Harper. 1996. A critical perspective on action research as a method for information systems research." *Journal of information Technology* 11.3, 235-246.

Peter Checkland and Sue Holwell. 1998. Action Research: Its Nature and Validity. *Systematic Practice and Action Research.* (Oct. 1997), 9–21.

Yvonne Dittrich. 2002. Doing Empirical Research on Software Development: Finding a Path between Understanding, Intervention, and Method Development. In *Social thinking—Software practice*. 243–262

Yvonne Dittrich, Kari Rönkkö, Jeanette Eriksson, Christina Hansson and Olle Lindeberg. 2008. Cooperative method development. *Empirical Software Engineering.* 13, 3, 231-260. DOI: 10.1007/s10664-007-9057-1

Kurt Lewin. 1947. Frontiers in Group Dynamics. *Human Relations* 1, 2 (1947), 143–153. DOI: 10.1177/001872674700100201

Lars Mathiassen. 1998. Reflective systems development. *Scandinavian Journal of Information Systems* 10, 1 (1998), 67–118

Lars Mathiassen. 2002. Collaborative practice research. *Information, Technology & People.* 15,4 (2002), 321–345

Lars Mathiassen, Mike Chiasson, and Matt Germonprez. 2012. Style Composition in Action Research Publication. *MIS quarterly. JSTOR* 36, 2 (2012), 347-363

Miroslaw Staron. Action research in software engineering: Metrics' research perspective. *International Conference on Current Trends in Theory and Practice of Informatics.* (2019), 39-49

Maung K. Sein, Ola Henfridsson, Sandeep Purao, Matti Rossi and Rikard Lindgren. 2011. Action design research. *MIS quarterly*. (2011), 37-56. DOI: 10.2307/23043488

Exemplars

Yvonne Dittrich, Kari Rönkkö, Jeanette Eriksson, Christina Hansson and Olle Lindeberg. 2008. Cooperative method development. *Empirical Software Engineering*. 13, 3 (Dec. 2007), 231-260. DOI: 10.1007/s10664-007-9057-1

Helle Damborg Frederiksen, Lars Mathiassen. 2005. Information-centric assessment of software metrics practices. IEEE Transactions on Engineering Eanagement. 52, 3 (2005), 350-362. DOI: 10.1109/TEM.2005.850737

Jakob Iversen and Lars Mathiassen. 2003. Cultivation and engineering of a software metrics program. Information Systems Journal. 13, 1 (2006), 3–19

Jakob Iversen. 1998. Problem diagnosis software process improvement. Larsen TJ, Levine L, DeGross JI (eds) Information systems: current issues and future changes.

Martin Kalenda, Petr Hyna, Bruno Rossi. *Scaling agile in large organizations: Practices, challenges, and success factors*. Journal of Software: Evolution and Process. Wiley Online Library 30, 10 (Oct. 2018), 1954 pages.

Miroslaw Ochodek, Regina Hebig, Wilhem Meding, Gert Frost, Miroslaw Staron. Recognizing lines of code violating company-specific coding guidelines using machine learning. *Empirical Software Engineering*. 25, 1 (Jan. 2020), 220-65.

Kari Rönkkö, Brita Kilander, Mats Hellman, Yvonne Dittrich. 2004. Personas is not applicable: local remedies interpreted in a wider context. In *Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices-Volume 1, Toronto, ON*, 112–120.

Thatiany Lima De Sousa, Elaine Venson, Rejane Maria da Costa Figueired, Ricardo Ajax Kosloski, and Luiz Carlos Miyadaira Ribeiro. Using Scrum in Outsourced Government projects: An Action Research. 2016. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, January 5, 2016, 5447-5456.

Hataichanok Unphon, Yvonne Dittrich. 2008. Organisation matters: how the organisation of software development influences the introduction of a product line architecture. In *Proc. IASTED Int. Conf. on Software Engineering*. 2008, 178-183

Case Study

"An empirical inquiry that investigates a contemporary phenomenon (the "case") in depth and within its real-world context, especially when the boundaries between phenomenon and context [are unclear]" (Yin 2017)

Application

This standard applies to empirical research that meets the following conditions.

- Presents a detailed account of a specific instance of a *phenomenon* at a *site*. The phenomenon can be virtually anything of interest (e.g. Unix, cohesion metrics, communication issues). The site can be a community, an organization, a team, a person, a process, an internet platform, etc.
- Features direct or indirect observation (e.g. interviews, focus groups)—see Lethbridge et al.'s (2005) taxonomy.
- Is not an experience report (cf. Perry et al. 2004) or a series of shallow inquiries at many different sites.

A case study can be brief (e.g. a week of observation) or longitudinal (if observation exceeds the natural rhythm of the site; e.g., observing a product over many releases). For our purposes, *case study* subsumes ethnography.

If data collection and analysis are interleaved, consider the **Grounded Theory Standard**. If the study mentions action research, or intervenes in the context, consider the **Action Research Standard.** If the study captures a large quantitative dataset with limited context, consider the **Exploratory Data Science Standard.**

Specific Attributes

Importance	Attribute
Essential	☐ explains why the case study approach is appropriate for the research question
	☐ justifies the selection of the case or site that was studied
	☐ describes the context of the case in rich detail
	☐ defines unit(s) of analysis
	☐ presents a clear and well-argued "chain of evidence" from observations to findings
	☐ clearly answers the research question(s)
Desirable	☐ reports the type of case study (see *Types of Case Studies*, below)
	☐ describes external events and other factors that may have affected the case or site
	☐ explains how researchers triangulated across data sources, informants or researchers
	☐ cross-checks interviewee statements (e.g. against direct observation or archival records)
	☐ uses quotations to *illustrate* findings (note: quotations should not be *the only* representation of a finding; each finding should be described independently of supporting quotations)
Extraordinary	☐ multiple, deep, fully-developed cases with cross-case triangulation
	☐ uses multiple judges and reports inter-rater reliability (cf. Gwet & Gwet 2002)
	☐ uses direct observation and clearly integrates direct observations into results
	☐ created a case study protocol beforehand and makes it publicly accessible

General Quality Criteria

Case studies should be evaluated using qualitative validity criteria such as credibility, multivocality, reflexivity, rigor and transferability (see **Glossary**). Quantitative quality criteria such as replicability, generalizability and objectivity typically do not apply.

Types of Case Studies

There is no standard way of conducting a case study. Case study research can adopt different philosophies, most notably (post-)positivism (Lee 1989) and interpretivism/constructivism (Walsham 1995), and serve different purposes, including:

- a **descriptive case study** describes—in vivid detail–a particular instance of a phenomenon
- an **emancipatory case study** identifies social, cultural, or political domination "that may hinder human ability" (Runeson and Host 2009), commensurate with a critical epistemological stance
- an **evaluative case study** evaluates a priori research questions, propositions, hypotheses or technological artifacts
- an **explanatory case study** explains how or why a phenomenon occurred, typically using a process or variance theory
- an **exploratory case study** explores a particular phenomenon to identify new questions, propositions or hypotheses
- an **historical case study** draws on archival data, for instance, software repositories
- a **revelatory case study** examines a hitherto unknown or unexplored phenomenon

Invalid Criticisms

- Does not present quantitative data; only collects a single data type.

- Sample of 1; findings not generalizable. The point of a case study is to study one thing deeply, not to generalize to a population. Case studies should lead to theoretical generalization; that is, concepts that are transferable in principle.
- Lack of internal validity. Internal validity only applies to explanatory case studies that seek to establish causality.
- Lack of reproducibility or a "replication package"; Data are not disclosed (qualitative data are often confidential).
- Insufficient number of length of interviews. There is no magic number; what matters is that there is enough data that the findings are credible, and the description is deep and rich.

Exemplars

Adam Alami, and Andrzej Wąsowski. 2019. Affiliated participation in open source communities. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 1-11

Michael Felderer and Rudolf Ramler. 2016. Risk orientation in software testing processes of small and medium enterprises: an exploratory and comparative study. *Software Quality Journal*. 24, 3 (2016), 519-548.

Audris Mockus, Roy T. Fielding, and James D. Herbsleb. 2002. Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology (TOSEM).* 11, 3 (2002), 309-346.

Helen Sharp and Hugh Robinson. 2004. An ethnographic study of XP practice. *Empirical Software Engineering.* 9, 4 (2004), 353-375.

Diomidis Spinellis and Paris C. Avgeriou. Evolution of the Unix System Architecture: An Exploratory Case Study. *IEEE Transactions on Software Engineering*. (2019).

Klaas-Jan Stol and Brian Fitzgerald. Two's company, three's a crowd: a case study of crowdsourcing software development. In *Proceedings of the 36th International Conference on Software Engineering*, 187–198, 2014.

Suggested Readings

Line Dube and Guy Pare. Rigor in information systems positivist case re-search: current practices, trends, and recommendations. 2003. *MIS quarterly.* JSTOR 27, 4 (Dec. 2003), 597–636. DOI: 10.2307/30036550

Shiva Ebneyamini, and Mohammad Reza Sadeghi Moghadam. 2018. Toward Developing a Framework for Conducting Case Study Research. *International Journal of Qualitative Methods.* 17, 1 (Dec. 2018)

Kilem Gwet. 2002. Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity. Statistical Methods for Inter-Rater Reliability Assessment Series, 2 (May 2002), 9 pages.

Barbara Kitchenham, Lesley Pickard, and Shari Lawrence Pfleeger. 1995. Case studies for method and tool evaluation. *IEEE software.* 12, 4 (1995), 52-62.

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.

Mathew Miles, A Michael Huberman and Saldana Johnny. 2014. *Qualitative data analysis: A methods sourcebook*.

Dewayne E. Perry, Susan Elliott Sim, and Steve M. Easterbrook. 2004. Case Studies for Software Engineers, In *Proceedings 26th International Conference on Software Engineering.* 28 May 2008, Edinburgh, UK, 736-738.

Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*. 14, 2, 131 pages.

Per Runeson, Martin Host, Austen Rainer, and Bjorn Regnell. 2012. Case study research in software engineering: Guidelines and examples. John Wiley & Sons.

Sarah J. Tracy. 2010. Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research. *Qualitative Inquiry. 16*, 10, 837–851. DOI: 10.1177/1077800410383121

Geoff Walsham, 1995. Interpretive case studies in IS research: nature and method. *European Journal of information systems.* 4,2, 74-81.

Robert K. Yin. 2017. *Case study research and applications: Design and methods*. Sage publications.

Grounded Theory

A study of a specific area of interest or phenomenon that involves iterative and interleaved rounds of qualitative data collection and analysis, leading to key patterns (e.g. concepts, categories)

Application

This standard applies to empirical inquiries that meet all of the following conditions:

- Explores a broad area of investigation without specific, up-front research questions.
- Applies theoretical sampling with iterative and interleaved rounds of data collection and analysis.
- Reports rich and nuanced findings, typically including verbatim quotes and samples of raw data.

For predominately qualitative inquiries that do not iterate between data collection and analysis or do not use theoretical sampling, consider the **Case Study Standard** or the **Qualitative Survey Standard**.

Specific Attributes

Importance	Attribute
Essential	☐ identifies the version of Grounded Theory used/adapted (Glaser, Strauss-Corbin, Charmaz, etc.)
	☐ explains how data source(s) were selected and accessed (e.g. participant sampling strategy)
	☐ explains how the research iterated between data collection and analysis using constant comparison and theoretical sampling
	☐ provides evidence of saturation; explains how saturation was achieved
	☐ explains how key patterns (e.g. categories) emerged from GT steps (e.g. selective coding)
	☐ provides clear chain of evidence from raw data (e.g. interviewee quotations) to derived codes, concepts, and categories
Desirable	☐ explains how and why study adapts or deviates from claimed GT version
	☐ presents a mature, fully-developed theory or taxonomy
	☐ includes highly diverse participants and/or data sources (e.g. software repositories, forums)
	☐ uses direct quotations extensively to support key points
	☐ explains how memo writing was used to drive the work
	☐ validates results (e.g. member checking, feedback from non-participant practitioners, research audits of coding with advisors/other researchers)
	☐ includes supplemental materials such as interview guide(s), coding schemes, coding examples, decision rules, or chain-of-evidence tables too large for main text
	☐ discusses transferability; characterizes the setting such that readers can assess transferability
	☐ compares results with (or integrates them into) prior theory or related research
	☐ explains theoretical sampling vis-à-vis the interplay between the sampling process, the emerging findings, and theoretical gaps perceived therein
	☐ reflects on how researcher's biases may have affected their analysis
	☐ explains the role of literature, especially where an extensive review preceded the GT study
Extraordinary	☐ triangulates with extensive quantitative data (e.g. questionnaires, sentiment analysis)
	☐ employs a team of researchers and explains their roles

Quality Criteria

Glaser, Strauss, Corbin and Charmaz advance inconsistent quality criteria. Using definitions in our **Glossary**, reviewers should consider common qualitative criteria such as **credibility**, **resonance**, **usefulness** and the degree to which results *extend* our cumulative knowledge. Quantitative quality criteria such as internal validity, construct validity, replicability, generalizability and reliability typically do not apply.

Examples of Acceptable Deviations

- In a study of sexual harassment at a named organization, detailed description of interviewees and direct quotations are omitted to protect participants.

Antipatterns

- Conducting data collection and data analysis sequentially; applying only analysis techniques of GT.
- Data analysis focusing on counting words, codes, concepts, or categories instead of interpreting.
- Presenting a tutorial on grounded theory instead of explaining how the current study was conducted.

- Small, heterogenous samples creating the illusion of convergence and theoretical saturation. For example, it is highly unlikely that a full theory can be derived only from interviews with 20 people.
- Focusing only on interviews without corroborating statements with other evidence (e.g. documents, observation).

Invalid Criticisms

- lack of quantitative data; causal analysis; objectivity, internal validity, reliability, or generalizability
- lack of replicability or reproducibility; not releasing transcripts
- lack of representativeness (e.g. of a study of Turkish programmers, 'how does this generalize to America?')
- research questions should have been different
- findings should have been presented as a different set of relationships, hypotheses, or a different theory.

Suggested Readings

Steve Adolph, Wendy Hall, and Philippe Kruchten. 2011. Using grounded theory to study the experience of software development. *Empirical Software Engineering*. 16, 4 (2011), 487-513.

Terry Rowlands, Neal Waddell, and Bernard McKenna. 2016. Are We There Yet? A Technique to Determine Theoretical Saturation. *Journal of Computer Information Systems*. 56, 1 (2016), 40-47.

Klaas-Jan Stol, Paul Ralph, and Brian Fitzgerald. 2016. Grounded theory in software engineering research: a critical review and guidelines. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. Association for Computing Machinery, New York, NY, USA, 120–131. DOI: 10.1145/2884781.2884833

Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sge publications.

Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

Exemplars

Barthélémy Dagenais and Martin P. Robillard. 2010. Creating and evolving developer documentation: understanding the decisions of open source contributors. In Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering (FSE '10). Association for Computing Machinery, New York, NY, USA, 127–136. DOI: 10.1145/1882291.1882312

Rashina Hoda, James Noble, and Stuart Marshall. 2012. Self-organizing roles on agile software development teams. *IEEE Transactions on Software Engineering*. IEEE 39, 3 (May 2012), 422-444. DOI: 10.1109/TSE.2012.30

Todd Sedano, Paul Ralph, and Cécile Péraire. 2017. Software development waste. In 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE). (May. 2017), 130-140. DOI: 10.1109/icse (2017).

Christoph Treude and Margaret-Anne Storey. 2011. Effective communication of software development knowledge through community portals. In *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering (ESEC/FSE '11)*. Association for Computing Machinery, New York, NY, USA, 91–101. DOI:10.1145/2025113.2025129

Michael Waterman, James Noble, and George Allan. 2015. How much up-front? A grounded theory of agile architecture. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. 1, (May 2015), 347-357.

Qualitative Surveys (Interview Studies)

A study comprising semi-structured or open-ended interviews

Application

This standard applies to empirical inquiries that meet all of the following criteria:

- Researcher(s) have synchronous conversations with one participant at a time
- Researchers ask, and participants answer, open-ended questions
- Participants' answers are recorded in some way
- Researchers apply some kind of qualitative data analysis to participants' answers

If researchers iterated between data collection and analysis, consider the **Grounded Theory Standard**. If respondents are all from the same organization, consider the **Case Study Standard**. If researchers collect written text or conversations (e.g. StackExchange threads), consider the **Discourse Analysis Standard**.

Specific Attributes

Importance	Attribute
Essential	☐ explains how interviewees were selected (i.e. sampling strategy; see **The Sampling Supplement**)
	☐ describes interviewees (e.g. demographics, work roles)
	☐ presents clear chain of evidence from interviewee quotations to proposed concepts
Desirable	☐ includes highly diverse participants
	☐ uses direct quotations extensively to support key points
	☐ EITHER: evaluates an a priori theory (or model, framework, taxonomy, etc.) using deductive coding with an a priori coding scheme based on the prior theory OR: synthesizes results into a new, mature, fully-developed and clearly articulated theory (or model, etc.) using some form of inductive coding (coding scheme generated from data)
	☐ validates results (e.g. using member checking)
	☐ EITHER uses audits (inductive coding) OR uses multiple coders and reports agreement statistics (deductive coding)
	☐ provides supplemental materials including interview guide(s), coding schemes, coding examples, decision rules, extended chain-of-evidence table(s)
	☐ discusses transferability; findings plausibly transferable to different contexts
	☐ compares results with (or integrates them into) prior theory or related research
	☐ reflects on how researchers' biases may have affected their analysis
Extraordinary	☐ employs multiple methods of data analysis (e.g. open coding vs. process coding; manual coding vs. automated sentiment analysis) with method-triangulation
	☐ employs longitudinal design (i.e. each interviewee participates multiple times) and analysis
	☐ employs probabilistic sampling strategy; statistical analysis of response bias

General Quality Criteria

An interview study should address appropriate qualitative quality criteria such as: **credibility, resonance**, **usefulness**, and **transferability** (see **Glossary**). Quantitative quality criteria such as internal validity, construct validity, generalizability and reliability typically do not apply.

Examples of Acceptable Deviations

- In a study of deaf software developers, the interviews are conducted via text messages.
- In a study of sexual harassment at named organizations, detailed description of interviewees and direct quotations are omitted to protect participants.
- In a study of barriers faced by gay developers, participants are all gay (but should be diverse on other dimensions).

Antipatterns

- Interviewing a small number of similar people, creating the illusion of convergence and saturation
- Mis-presenting a qualitative survey as grounded theory or a case study.

Invalid Criticisms

- Lack of quantitative data; causal analysis; objectivity, internal validity, reliability, or generalizability.
- Lack of replicability or reproducibility; not releasing transcripts.

- Lack of probability sampling, statistical generalizability or representativeness unless representative sampling was an explicit goal of the study.
- Failure to apply grounded theory or case study practices. A qualitative survey is not grounded theory or a case study.

Notes

- A qualitative survey generally has more interviews than a case study that triangulates across different kinds of data.

Suggested Readings

Michael Quinn Patton. 2002. *Qualitative Research and Evaluation Methods*. 3rd ed. Sage Publications.
Herbert J. Rubin and Irene S. Rubin. 2011. *Qualitative interviewing: The art of hearing data*. Sage.
Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.

Exemplars

Marian Petre. 2013. UML in practice. In *Proceedings of the 35th International Conference on Software Engineering,* San Francisco, USA, 722-731.

Paul Ralph and Paul Kelly. 2014. The dimensions of software engineering success. In *Proceedings of the 36th International Conference on Software Engineering (ICSE 2014)*. Association for Computing Machinery, New York, NY, USA, 24–35. DOI: 10.1145/2568225.2568261

Paul Ralph and Ewan Tempero. 2016. Characteristics of decision-making during coding. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering (EASE '16).* Association for Computing Machinery, New York, NY, USA, Article 34, 1–10. DOI:10.1145/2915970.2915990

Quantitative Methods

Experiments (with Human Participants)

A study in which an intervention is deliberately introduced to observe its effects on some aspects of reality under controlled conditions

Application

This standard applies to controlled experiments and quasi-experiments that meet all of the following conditions:

- manipulates one or more independent variables
- controls many extraneous variables
- applies each treatment independently to several experimental units
- involves human participants

In true experiments, experimental units are randomly allocated across treatments; quasi-experiments lack random assignment. Experiments include between-subjects, within-subjects and repeated measures designs. For experiments without human participants, see the **Exploratory Data Science Standard** or the **Engineering Research Standard**.

Specific Attributes

Importance	Attribute
Essential	☐ describes how characteristics of phenomenon under investigation relate to experimental constructs
	☐ states formal hypotheses
	☐ justifies use of one-sided hypotheses (if any) based on face validity or previous work
	☐ describes independent, dependent and extraneous variables; how extraneous vars are controlled
	☐ describes the research design and protocol including *treatments, materials*, *tasks*, *design* (e.g. 2x2 factorial), *participant allocation*, *period and sequences* (for crossover designs), and logistics
	☐ design and protocol *appropriate* (not optimal) for stated research questions and hypotheses
	☐ EITHER: uses random assignment; explains logistics (e.g. how random numbers were generated) OR: justifies why random assignment is impractical or unethical (compelling reason needed); and mitigates unequal groups threat to validity (e.g. using pre-test/post-test and matched subjects design)
	☐ describes experimental objects (e.g. real or toy system) and their characteristics (e.g. size, type);
	☐ justifies selection of experimental objects; checks for object-treatment confounds[8]
	☐ describes and justifies how the dependent variable is measured (including units, instruments)
	☐ describes how independent and dependent variables are measured
	☐ describes participants (e.g. age, gender, education, relevant experience or preferences)
	☐ reports distribution-appropriate descriptive and inferential statistics; enumerates and checks assumptions[9]; justifies tests used
	☐ reports effects sizes with confidence intervals (if using frequentist approach)
	☐ EITHER: shares raw, de-identified data OR: explains why sharing raw data is impractical or unethical
	☐ discusses construct, conclusion internal and external validity
	☐ discusses alternative interpretations of results
Desirable	☐ justifies hypotheses and Bayesian priors (if applicable) based on previous studies and theory
	☐ discusses alternative experimental designs and why they were not used (e.g. validity trade-offs)
	☐ includes visualizations of data distributions
	☐ cites statistics papers to support any nuanced issues or unusual approaches
	☐ explains deviations between design and execution, and their implications[10]
	☐ includes supplementary material: complete, algorithmic research protocol, task materials, de-identified dataset, analyses scripts
	☐ named experiment design (e.g. simple 2-group, 2x2 factorial, randomized block)
	☐ presents a-priori power analysis and sufficient *n* for expected effect sizes.
	☐ analyzes construct validity of dependent variable
	☐ uses and reports manipulation checks
	☐ pre-registration of hypotheses and design where venue allows

[8] e.g., in an experiment where control group applies Test-Last (TL) with Object 1 while treatment group applies Test-Driven-Development (TDD) with Object 2, the experimental object is confounded with the treatment.
[9] visual methods of checking assumptions are often as good as or better than statistical tests
[10] e.g. dropouts affecting balance between treatment and control group

Extraordinary	☐	reports multiple experiments or replications in different cultures or regions
	☐	uses multiple methods of data collection; data triangulation
	☐	longitudinal data collection with appropriate time-series analysis

General Quality Criteria

Conclusion validity, construct validity, internal validity, reliability, objectivity, reproducibility

Invalid criticisms

- participants are students—appropriateness of participant characteristics should be judged based on the context, desired level of control, trade-off choices between internal and external validity, and the specifics of the technology (i.e. method, technique, tool, process, etc.) under evaluation; the choice must be explained in the paper
- low external validity
- the experiment is a replication
- the reviewer would have investigated the topic in any other way than an experiment

Antipatterns

- using bad proxies for dependent variables (e.g. task completion time as a proxy for task complexity)
- quasi-experiments without a good reason[11]
- treatments or response variables are poorly described
- inappropriate design for the conditions under which the experiment took place
- data analysis technique used does not correspond to the design chosen or data characteristics (e.g. using an independent samples t-test on paired data)
- validity threats are simply listed without linking them to results
- hypotheses are missing

Suggested Reading

Nathaniel L. Gage and Julian C. Stanley. 1963. Experimental and Quasi-experimental Designs For Research. Chicago: R. McNally.

Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl. 2008. Reporting Experiments in Software Engineering. *Guide to Advanced Empirical Software Engineering*. 201-228.

Natalia Juristo and Ana M. Moreno. 2001. *Basics of Software Engineering Experimentation*. Springer Science & Business Media.

Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media.

Martín Solari, Sira Vegas, and Natalia Juristo. 2018. Content and structure of laboratory packages for software engineering experiments. *Information and Software Technology.* 97, 64-79.

Sira Vegas, Cecilia Apa, and Natalia Juristo. 2015. Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering*. IEEE 42, 2 (2015), 120-135.

Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag IK Sjøberg. 2009. A systematic review of quasi-experiments in software engineering. *Information and Software Technology*. 51, 1 (2009), 71-82.

Davide Falessi, Natalia Juristo, Claes Wohlin, Burak Turhan, Jürgen Münch, Andreas Jedlitschka, and Markku Oivo, Empirical Software Engineering Experts on the Use of Students and Professionals in Experiments, *Empirical Software Engineering.* 23, 1 (2018), 452-489.

Robert Feldt, Thomas Zimmermann, Gunnar R. Bergersen, Davide Falessi, Andreas Jedlitschka, Natalia Juristo, Jürgen Münch et al. 2018. Four commentaries on the use of students and professionals in empirical software engineering experiments. *Empirical Software Engineering*. 23, 6 (Nov. 2018), 3801-3820.

Kitchenham, Barbara, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong. 2017. Robust statistical methods for empirical software engineering. *Empirical Software Engineering*. 22, 2 (2018), 579-630.

Andreas Zeller, Thomas Zimmermann, and Christian Bird. 2011. Failure is a four-letter word: a parody in empirical research. In *Proceedings of the 7th International Conference on Predictive Models in Software Engineering (Promise '11).* Association for Computing Machinery, New York, NY, USA, Article 5, 1–7. DOI: 10.1145/2020390.2020395

Exemplars

Dag IK Sjøberg, Aiko Yamashita, Bente CD Anda, Audris Mockus, and Tore Dybå. 2012. Quantifying the Effect of Code Smells on Maintenance Effort. *IEEE Transactions on Software Engineering*. 39, 8 (Dec. 2012), 1144-1156. DOI: 10.1109/TSE.2012.89.

[11] Quasi-experiments are appropriate for pilot studies or when assignment is beyond the researcher's control (e.g. assigning students to two different sections of a course). Simply claiming that a study is "exploratory" is not sufficient justification.

Ayse Tosun, Oscar Dieste, Davide Fucci, Sira Vegas, Burak Turhan, Hakan Erdogmus, Adrian Santos et al. 2017. An industry experiment on the effects of test-driven development on external quality and productivity. *Empirical Software Engineering*. *22*, 6 (Dec. 2016), 2763-2805.

Kai Petersen, Kari Rönkkö, and Claes Wohlin. 2008. The impact of time controlled reading on software inspection effectiveness and efficiency: a controlled experiment. In Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement (ESEM '08). Association for Computing Machinery, New York, NY, USA, 139–148. DOI:10.1145/1414004.1414029

Eduard P. Enoiu, Adnan Cauevic, Daniel Sundmark, and Paul Pettersson. 2016. A controlled experiment in testing of safety-critical embedded software. In *2016 IEEE International Conference on Software Testing, Verification and Validation (ICST),* 11-15 April, Chicago, IL, USA. IEEE. 1-11.

Yang Wang and Stefan Wagner. 2018. Combining STPA and BDD for safety analysis and verification in agile development. In Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings (ICSE '18). Association for Computing Machinery, New York, NY, USA, 286–287. DOI:10.1145/3183440.3194973

Evrim Itir Karac, Burak Turhan, and Natalia Juristo. 2019. A Controlled Experiment with Novice Developers on the Impact of Task Description Granularity on Software Quality in Test-Driven Development. *IEEE Transactions on Software Engineering.*

Questionnaire Surveys

A study in which a sample of respondents answer a series of questions. Questions are typically answered through a computerized or paper form and mostly structured

Application

This guideline applies to studies in which:

- a sample of participants answer, predefined, mostly closed-ended questions (typically online or on paper)
- researchers systematically analyze participants' answers

Surveys can be descriptive, exploratory or confirmatory. Confirmatory surveys can test individual propositions or complex theories. This standard does not apply to questionnaires comprising predominately open-ended questions[12], literature surveys (see the **Systematic Review Standard**), longitudinal or repeated measures studies (see the **Longitudinal Studies Standard**), or the demographic questionnaires typically given to participants in controlled experiments (see the **Experiments Standard**).

Specific Attributes

Section	Attribute
Essential	☐ identifies the target population & defines the sampling strategy (see the **Sampling Supplement**)
	☐ provides questionnaire instrument (e.g. as supplemental file)
	☐ EITHER: provides study artifacts; i.e., instrument(s), code books, analysis scripts and dataset(s) (addressing potential anonymity and confidentiality issues) OR: describes in detail study artifacts and justifies why they are not provided
	☐ the questionnaire design matches the research aims (i.e. questions are mapped to research objectives) and the target population (wording and format of the questions)
	☐ describes how participants were selected, including invitations and incentives
	☐ step-by-step, systematic, replicable description of data collection and analysis
	☐ describes how responses were managed/monitored, including contingency actions for non-responses and drop-outs
	☐ EITHER: measures constructs using (or adapting) validated scales OR: analyzes construct validity (e.g. content, convergent, discriminant, predictive) ex post
	☐ explains handling of missing data (e.g. imputation, weighting adjustments, discarding)
	☐ acknowledges generalizability threats; discusses how respondents may differ from target population
	☐ analyzes response rates
Desirable	☐ characterizes the target population including demographic information (e.g. culture, knowledge)
	☐ defines and estimates the size of the population strata (if applicable)
	☐ accounts for the principles of research ethics (e.g. informed consent, re-identification risk)
	☐ explains and justifies instrument design and choice of scales (e.g. by research objectives or by analogy to similar studies).
	☐ validates whether the items, layout, duration, and technology are appropriate (e.g. using pilots, test-retest, or expert and non-expert reviews).
	☐ reports how the instrument has evolved through the validation process (if at all)
	☐ applies techniques for improving response rates (e.g. incentives, reminders, targeted advertising)
	☐ analyzes response bias (quantitatively)
	☐ discusses possible effect of incentives (e.g. on voluntariness, response rates, response bias) if used
	☐ describes the stratification of the analysis (if stratified sampling is used)
	☐ clearly distinguishes evidence-based results from interpretations and speculation[13]
Extraordinary	☐ provides feasibility check of the anticipated data analysis techniques
	☐ reports on the scale validation in terms of dimensionality, reliability, and validity of measures

[12] There is currently no standard for predominately open-ended questionnaire surveys. One exemplar readers could draw from is: Daniel Graziotin, Fabian Fagerholm, Xiaofeng Wang, and Pekka Abrahamsson. 2018. "What happens when software developers are (un)happy." *Journal of Systems and Software* 140, 32-47.

[13] Simply separating results and discussion into different sections is typically sufficient. No speculation in the results section.

General Quality Criteria

Survey studies should address quantitative quality criteria such as **internal validity**, **construct validity**, **external validity**, **reliability** and **objectivity** (see **Glossary**).

Variations

- **Descriptive surveys** provide a detailed account of the properties of a phenomenon or population.
- **Exploratory surveys** generate insights, hypotheses or models for further research.
- **Confirmatory surveys** testing formal (e.g. causal) propositions to explain a phenomenon.

Invalid Criticism

- Not reporting response rate for open public subscription surveys (i.e. surveys open to the anonymous public so that everyone with a link—typically broadcasted among social networks—can participate).
- Failure to release full data sets despite the data being sensitive.
- Claiming the sample size is too small without justifying why the sample size is insufficient to answer the research questions.
- Criticizing the relevance of a survey on the basis that responses only capture general people's perceptions.
- The results are considered hardly surprising or controversial.
- The results do not accord with the reviewer's personal experience or previous studies.

Suggested Readings

Don Dillman, Jolene Smyth, and Leah Christian. 2014. Internet, phone, mail, and mixed-mode surveys: the tailored design method. John Wiley & Sons.

Mark Kasunic. 2005. Designing an effective survey. Carnegie-Mellon Univ Pittsburgh PA Software Engineering Inst.

Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. *An empirically evaluated checklist for surveys in software engineering.* Information and Software Technology. 119 (2020).

Stefan Wagner, Daniel Mendez, Michael Felderer, Daniel Graziotin, Marcos Kalinowski. Challenges in Survey Research. In: Contemporary Empirical Methods in Software Engineering, *Springer,* 2020.

Paul Ralph and Ewan Tempero. 2018. Construct Validity in Software Engineering Research and Software Metrics. In Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018 (EASE'18). Association for Computing Machinery, New York, NY, USA, 13–23. DOI:10.1145/3210459.3210461

Marco Torchiano, Daniel Méndez, Guilherme Horta Travassos, and Rafael Maiani de Mello. 2017. Lessons learnt in conducting survey research. In *Proceedings of the 5th International Workshop on Conducting Empirical Studies in Industry (CESI '17)*, 33–39. DOI:10.1109/CESI.2017.5

Torchiano Marco and Filippo Ricca. Six reasons for rejecting an industrial survey paper. In *2013 1st International Workshop on Conducting Empirical Studies in Industry (CESI).* (2013), 21-26.

Exemplars

Stefan Wagner, Daniel Méndez Fernández, Michael Felderer, Antonio Vetrò, Marcos Kalinowski, Roel Wieringa, Dietmar Pfahl, Tayana Conte, Marie-Therese Christiansson, Desmond Greer, Casper Lassenius, Tomi Männistö, Maleknaz Nayebi, Markku Oivo, Birgit Penzenstadler, Rafael Prikladnicki, Guenther Ruhe, André Schekelmann, Sagar Sen, Rodrigo Spínola, Ahmed Tuzcu, Jose Luis De La Vara, and Dietmar Winkler. 2019. *Status Quo in Requirements Engineering: A Theory and a Global Family of Surveys.* ACM Trans. Softw. Eng. Methodol. 28, 2, Article 9 (April 2019), 48 pages. DOI:10.1145/3306607

D. Méndez Fernández, Stefan Wagner, Marcos Kalinowski, Michael Felderer, Priscilla Mafra, Antonio Vetrò, Tayana Conte et al. Naming the Pain in Requirements Engineering: Contemporary Problems, Causes, and Effects in Practice. In *Empirical software engineering*. 22, 5 (2016), 2298—2338.

Jingyue Li, Reidar Conradi, Odd Petter Slyngstad, Marco Torchiano, Maurizio Morisio, and Christian Bunse. A State-of-the-Practice Survey on Risk Management in Development with Off-The-Shelf Software Components. In *IEEE Transactions on Software Engineering*. 34, 2 (2008), 271-286.

Systematic Reviews

A study that appraises, analyses, and synthesizes primary or secondary literature to provide a complete, exhaustive summary of current evidence regarding one or more specific topics or research questions

Application

- Applies to studies that systematically find and analyze existing literature about a specified topic
- Applies both to secondary and tertiary studies
- Does not apply to ad-hoc literature reviews, case surveys or advanced qualitative synthesis methods (e.g. meta-ethnography)

Specific Attributes

Section	Attribute
Essential	☐ step-by-step, systematic, replicable description of search process including search terms[14]
	☐ defines clear inclusion and exclusion criteria
	☐ specifies the data extracted from each primary study[15]; explains relationships to research questions
	☐ describes in detail how data were extracted and synthesized (can be qualitative or quantitative)
	☐ describes coding scheme(s) and their use
	☐ clear chain of evidence from the extracted data to the answers to the research question(s)
	☐ presents conclusions or recommendations for practitioners/non-specialists
	☐ identifies method (e.g. systematic review, meta-analysis, mapping study, narrative synthesis, etc.)
Desirable	☐ provides replication package including protocol, search terms, search results, selection process results; complete dataset, analysis scripts; examples of coding, decision rules or edge cases
	☐ mitigates sampling bias and publication bias, using some combination of: (i) manual and keyword automated searches; (ii) backward and forward snowballing searches; (iii) checking profiles of prolific authors in the area; (iv) searching both formal databases (e.g. ACM Digital Library) and indexes (e.g. Google Scholar); (v) searching for relevant dissertations; (vi) searching pre-print servers (e.g. arXiv); (iiv) soliciting unpublished manuscripts through appropriate listservs or social media; (iiiv) contacting known authors in the area.
	☐ demonstrates that the search process is sufficiently rigorous for the systematic review goals[16]
	☐ assesses quality of primary studies; explains how quality was assessed
	☐ assesses coverage using funnel plots or percentage of known papers found
	☐ (positivist reviews), uses 2+ independent analysts; analyzes inter-rater reliability (e.g. KALPHA)
	☐ (interpretivist reviews) reflects on how researcher's biases may have affected their analysis
	☐ consolidates results using tables, diagrams, or charts; PRISMA flow diagram (cf. Moher et al. 2009)
	☐ performs analysis through an existing or new conceptual framework (qualitative synthesis)
	☐ uses meta-analysis methods appropriate for primary studies; does not use vote counting
	☐ integrates results into prior theory or research; identifies gaps, biases, or future directions
	☐ presents results as practical, evidence-based guidelines for practitioners, researchers, or educators
Extraordinary	☐ two or more researchers independently undertaking the preliminary search process before finalizing the search scope and search keywords
	☐ contacted primary study authors to ensure interpretations were correct, and elicit additional details not found in the papers such as access to raw data

Examples of Acceptable Deviations

- No attempts to mitigate publication bias in a study explicitly examining a specific venue's (e.g. CACM or ICSE) coverage of a given topic.
- Using probability sampling on primary studies when there are too many to analyze (i.e. thousands).
- No recommendations for practitioners in a study of a methodological issue (e.g. representative sampling).

Anti-Patterns

- A laundry-list description of the studies (A found X, B found Y, …), rather than a synthesis of the findings.

[14] Searches can be manual or automated or a combination of both

[15] Primary studies are the studies that are being reviewed. In a tertiary study, the "primary studies" are themselves reviews.

[16] e.g. formal meta-analysis of experiments has higher requirements for completeness than mapping studies of broad topic areas

- Relying on characteristics of the publication venues as a proxy for the quality of the primary studies instead of assessing primary studies' quality explicitly.
- Reviewing an area in which there are too few high-quality primary studies to draw reliable conclusions.

Suggested Readings

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6, 7: e1000097. doi:10.1371/journal.pmed1000097

Michael Borenstein and Larry V. Hedges and Julian P.T. Higgins and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis.* John Wiley & Sons Ltd.

Daniela S. Cruzes and Tore Dybå. 2010. Synthesizing evidence in software engineering research. In Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '10). Association for Computing Machinery, New York, NY, USA, Article 1, 1–10. DOI:10.1145/1852786.1852788

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering.

Matthew B. Miles and A. Michael Huberman and Jonny Saldana. 2014. Qualitative Data Analysis: A Methods Sourcebook. Sage Publications Inc.

Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE).* (Jun. 2008), 1-10.

Supplements

Information Visualization

Diagrams that map quantitative values to visual objects their visual attributes to aid understanding

Application

This crosscutting guideline applies to visualizations of quantitative data. Graphical visualization consists of diagrams that map quantitative values to visual objects (e.g. segments, points, circles) and to their visual attributes (e.g. length, position, size, color). Their goal is to support the reader in a data understanding task. It does not apply to:

- Software visualization, such as showing structure or architecture of a software product.
- Diagrams not encoding quantitative values, e.g. UML activity diagrams, BPMN diagrams, flow charts.
- Tables that encode data in textual format.

Attributes

☐ proportionality

- the values/measures are reported in a uniformly proportional way: the ratio of two values is equal to the geometrical ratio on paper (or screen) of the corresponding visual attributes (length, area, slope, etc.) [lie factor]
- the visual attributes provide an accurate perception of the proportion, according to the attribute ranking:
  - position along a common scale
  - position along identical scales
  - length
  - angle/slope
  - area
- diagrams are rendered in 2D and refrain from 3D perspectives that might alter the perception of dimensions
- encoding of ordinal measures through colors use saturation and lightness and avoid rainbow palettes

☐ utility

- all the elements in the diagram convey useful information or support clarity
- the diagram does not contain chartjunk or over-designed elements that interfere with perception or understanding
- the background is light and uniform
- there are no decorative 3D effects
- bright and saturated colors are used solely for emphasis
- grids are light and do not obscure data
- annotations are less prominent than data

☐ clarity

- the diagram layout and text annotation support the understanding of the data and make the visualization as much self-contained as possible
  - the title (or figure caption) concisely conveys what is the content of the visualization and the intended message
  - text annotations guide the reader in understanding the message
- direct labeling is used instead of a separate legend especially when there are more than two color codes
- color encoding of categorical measures is limited to at most 5 distinct levels; more colors are too difficult to discriminate
- when data points are very dense, appropriate techniques are applied to mitigate overplotting
- axes and the relative tick marks are labelled
- the size of text is large enough to make it readable.
- image format is preferably vectorial, if a raster format is used it must have sufficient resolution.

☐ diagram design

- the diagram contains at most two axes unless a surface (function of two variables) is the standard representation
- use of logarithmic scales is explicitly highlighted
- data objects (e.g. bars) are sorted in a meaningful way (e.g. ascending, descending or grouped) to ease comparison
- non-interactive visualization should serve a single understanding task
- the type of visualization is appropriate for the visual understanding task that it is intended to support (Table 1)
- *(optional)* several data series should usually be reported using multiple small diagrams rather than in a single crowded diagram

Table 1: When to use which visualization

Understanding task	Commonly used type of visualization
Comparison	Bar plot, Dot plot, Heatmap, Strip plot, Isotype
Correlation	Scatter plot, Bubble chart, Slope chart, Dumbbell plot, Diverging bars
Deviation	Bullet graph, Gauge
Distribution	Histogram, Frequency polygon, Cumulative density, Quantile-quantile plot, Boxplot, Violin plot
Geo location	Choroplet map, Cartogram
Part-to-whole	Bar plot, Stacked bars, Treemap, Waffle, Pie
Ranking	Bar plot, Dot plot, Lollypop
Time series	Line plot, Bar plot, Streamgraph

Anti-patterns

- using truncated bars to exaggerate differences, compromising proportionality instead of using other representations (e.g. dot plots)
- using pie charts for more than 5 categories and/or without direct labelling of the slices
- using dual vertical scales that are difficult to read and lend themselves to ambiguity due to the arbitrary selection of axis ranges
- using any 3D effect or decoration that may alter perception

Suggested Readings

Colin Ware. 2000. Information Visualization: Perception for Design. Morgan Kaufmann Publishers, Inc., San Francisco, California.

David Borland and R. M. Taylor Ii. 2007. Rainbow Color Map (Still) Considered Harmful. *IEEE Computer Graphics and Applications*. 27, 2, 14- 17.

Financial Times Visual Journalism Team. Visual Vocabulary. Retrieved July 12, 2020 from https://ft.com/vocabulary

Claus O. Wilke. Fundamentals of Data Visualization, O'Reilly, 2019. Retrieved July 12, 2020 from https://serialmentor.com/dataviz/

Simon Fear. Publication quality tables in LATEX. 2020. Retrieved July 12, 2020 from http://mirrors.ctan.org/macros/latex/contrib/booktabs/booktabs.pdf

Registered Reports

An empirical study that is published in two phases: the plan (RR1) and the results of executing the plan (RR2)

Definition

"Registered reports" refers to studies that are conducted in two phases:

- researchers create and publish a study plan, or phase 1 registered report (RR1), which is accepted *in principle* by a publishing venue;
- researchers execute the plan and write up the results, or a phase 2 registered report (RR2), which then receives *final acceptance* from the same publishing venue.

Pre-registration refers to depositing the RR1 somewhere publicly visible. Pre-registration aims to prevent hypothesising after results are known, to mitigate unconscious researcher bias in data analysis, and combat publication bias.

Application

This standard applies to positivist, confirmatory (i.e. hypothesis-testing) studies with tightly-scoped analysis approaches. Pre-registering interpretivist, qualitative or exploratory research remains controversial.

Specific Attributes (RR1)

Section	Attribute
Essential	☐ meets all essential criteria in *The General Standard* except those that require data: o does not present results o does not validate assumptions of statistical tests o does not discuss implications o does not contribute to collective body of knowledge o does not support conclusions with evidence or arguments
	☐ meets all essential criteria, in applicable empirical standards, that can be met before data collection[17]
	☐ justifies importance of the purpose, problem, objective, or research question(s)
	☐ describes the research method in detail sufficient for an independent researcher to exactly replicate the proposed data collection and analysis procedures
	☐ the stated hypotheses can be tested with the data the researchers propose to collect
Desirable	☐ presents preliminary data (e.g. from a pilot study) to justify the chosen approach (e.g. probability distributions).
	☐ includes a conditional structure (e.g. pre-specifying different tests for normal and non-normal distributions)
	☐ explains how the study will change based on the results of data analysis (i.e. conditional analysis)[18]

Specific Attributes (RR2)[19]

Section	Attribute
Essential	☐ meets all essential criteria in *The General Standard* (no exceptions)
	☐ introduction, rationale and stated hypotheses are the same as the approved RR1 submission except for improvements based on feedback from RR1 reviews
	☐ EITHER: adheres precisely to the registered procedures OR: thoroughly justifies all deviations and explains how they affect the final analysis.
	☐ deviations, if any, are *not* justified based on the data
	☐ clearly designates as exploratory any unregistered post hoc analyses
	☐ unregistered post hoc analysis, if any, are justified, methodologically sound, and informative
Desirable	☐ provides evidence that data was collected after RR1 plan is accepted
Extraordinary	☐ generates novel insights into the concept, process benefits, or limitations of registered reports

[17] e.g. presents power analysis; describes how card sorting will be executed, lists anticipated statistical tests

[18] e.g. as a decision tree; while not strictly required, omitting conditional analysis is extraordinarily risky for the authors

[19] Adapted from https://osf.io/pukzy/ by CC-BY

Antipatterns

- deviating from the RR1 plan because it constrains exploratory research, when the plan did not mention exploration
- postdictive deviations from RR1 plan; i.e., changes made knowing how they would affect the outcome of the study
- Pre-registrations that are not verifiably committed prior to data collection, e.g. not time-stamped.

Invalid Criticisms

- RR1: Insisting on complete data collection or detailed analysis and results
- RR1: Rejecting exploratory or qualitative research: all kinds of research can be pre-registered even it's not covered here
- RR2: in hindsight, the RR1 plan was not appropriate (RR2 reviews should not criticize any aspect of the RR1 plan)
- RR2: results are not statistically significant, novel, relevant or compelling; effect sizes too small

Justifying Deviations

Reviewers should not expect research to go exactly according to plan or authors to foresee every possible problem. Changes are acceptable as long as they are justified and not *postdictive* (i.e. changes made knowing how they will affect results). For example, researchers might drop a mistranslated question in a multi-lingual questionnaire survey.

Suggested Readings

Center for Open Science. Future-proof your research. Preregister your next study. Retrieved July 12, 2020 from https://www.cos.io/our-services/prereg

Center for Open Science. Registered reports: Peer review before results are known to align scientific values and practices. Retrieved July 12, 2020 from https://cos.io/rr/

Center for Open Science. Template reviewer and author guidelines. Retrieved July 12, 2020 from https://osf.io/pukzy/

Ben Goldacre, Nicholas J DeVito, Carl Heneghan, Francis Irving, Seb Bacon, Jessica Fleminger and Helen Curtis. 2018. Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. BMJ 2018;362:k3218 DOI: 10.1136/bmj.k3218

Wiseman R, Watt C, Kornbrot D. 2019. Registered reports: an early example and analysis. PeerJ 7:e6232 10.7717/peerj.6232

Methodological Guidelines and Meta-Science

A paper that analyses an issue of research methodology or makes recommendations for conducting research

Application

This standard applies to papers that provide analysis of one or more methodological issues, or advice concerning some aspect of research.

- may or may not include primary or secondary empirical data or analysis.
- may consider philosophical or practical issues
- may simultaneously be a methodology paper and an empirical study, to which another standard also applies; for example, if a paper reports a case study and then gives advice about a methodological issue illuminated by the case study, consider both this standard and the **Case Study Standard**.

Specific Attributes

Section	Attribute
Essential	☐ presents information that is useful for other researchers ☐ presents clear, valid arguments supporting recommendations
Desirable	☐ synthesizes related work from reference disciplines ☐ provides insight specifically for software engineering; goes beyond summarizing methodological guidance from existing works or reference disciplines; ☐ results integrated back into prior theory or research ☐ develops helpful artifacts (e.g. checklists, templates, tests, tools, sets of criteria)
Extraordinary	☐ includes an empirical study (e.g. a systematic literature review) that motivates the analysis of guidance ☐ quantitative simulation illustrating methodological issues

General Criteria

- **comprehensiveness** of analysis or guidance provided
- **usefulness** to the research community
- quality of **argumentation** supporting analysis of guidance
- degree of **integration** with previous work, both in software engineering and in reference disciplines

Antipatterns

- overreaching; informal logical fallacies (e.g. straw man argument, appeal to popularity, shifting the burden of proof)[20]
- discussing an issue without clear conclusions; failing to provide clear guidelines
- attacking individual studies or researchers; hypothetical examples should be used to avoid engendering animosity

Invalid Criticisms

- Guidelines are not based on empirical evidence. Empirically testing meta-scientific propositions is typically impractical or impossible. Reviewers should evaluate the face validity, comprehensiveness and usefulness of the guidelines. It is not appropriate to reject methodological guidelines over lack of empirical support.

Notes

Because metascientific claims cannot be justified empirically:

- reviewers of methodology papers must themselves be experts in the methodology, so that they can evaluate the reasonableness of the guidelines
- reviewers should be more pedantic in critiquing the discussion and guidelines than for an empirical paper

Exemplars

Natalia Juristo and Ana M. Moreno. 2001. *Basics of Software Engineering Experimentation.* Springer Science & Business Media.
Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering.
Barbara Kitchenham, Lesley Pickard, and Shari Lawrence Pfleeger. 1995. Case studies for method and tool evaluation. *IEEE software.* 12, 4 (1995), 52-62.

[20] citing seminal works is not the "appeal to authority" fallacy

Paul Ralph. 2019. Toward methodological guidelines for process theories and taxonomies in software engineering. *IEEE Transactions on Software Engineering*. 45, 7 (Jan. 2018), 712-735.

Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering.* 14, 2, article 131.

Miroslaw Staron. 2019. Action Research in Software Engineering: Theory and Applications. In *International Conference on Current Trends in Theory and Practice of Informatics*. 39-49.

Klaas-Jan Stol, Paul Ralph, and Brian Fitzgerald. 2016. Grounded theory in software engineering research: a critical review and guidelines. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. Association for Computing Machinery, New York, NY, USA, 120–131. DOI:10.1145/2884781.2884833

Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media.

Helen Sharp, Yvonne Dittrich, and Cleidson RB De Souza. 2016. "The role of ethnographic studies in empirical software engineering." *IEEE Transactions on Software Engineering* 42, 8: 786-804.

Yvonne Dittrich, Kari Rönkkö, Jeanette Eriksson, Christina Hansson, & Olle Lindeberg (2008). Cooperative method development. *Empirical Software Engineering,* 13, 3, 231–260.

Open Science

The practice of maximizing the accessibility and transparency of science

Application

The open science supplement applies to all research.

Principle

Artifacts related to a study and the paper itself should, in principle, be made available on the Internet:

- without any barrier (e.g. paywalls, registration forms, request mechanisms),
- under an appropriate open license that specifies purposes for re-use and re-purposing,
- properly archived and preserved,

provided that there are no ethical, legal, technical, economical, or practical barriers preventing their disclosure.

Specific Attributes

Section	Attribute
Desirable	☐ includes a section named *data availability* (typically after conclusion) ☐ EITHER: links to supplementary materials OR explains why materials cannot be released (reasons for limited disclosure of data should be trusted) ☐ includes supplementary materials such as: raw, deidentified or transformed data, extended proofs, analysis scripts, software, virtual machines and containers, or qualitative codebooks. ☐ archives supplementary materials on preserved digital repositories such as zenodo.org, figshare.com, softwareheritage.org, osf.io, or institutional repositories ☐ releases supplementary material under a clearly-identified open license such as CC0 or CC-BY 4.0

General Criteria

Rather than evaluating reproducibility or replicability in principle, reviewers should focus on the extent to which artifacts that can be released, are released.

Invalid Criticisms

Researchers should not complain that a study involves artifacts which— for good reasons—cannot be released.

Examples of Acceptable Deviations

- dataset is not released because it cannot be safely deidentified (e.g. interview transcripts; videos of participants)
- source code is not released because it is closed-source and belongs to industry partner

Notes

- authors are encouraged to self-archive their pre- and post-prints in open and preserved repositories
- open science is challenging for qualitative studies; reviewers should welcome qualitative studies which open their artifacts even in a limited way
- personal or institutional websites, version control systems (e.g. GitHub), consumer cloud storage (e.g. Dropbox), and commercial paper repositories (e.g. ResearchGate; Academia.edu) do not offer properly archived and preserved data.

Suggested Readings

Daniel Graziotin. 2020. Open science policies. Retrieved July 12, 2020 from https://ineed.coffee/open-science-policies/

Daniel Graziotin. 2020. SIGSOFT open science policies. Retrieved July 12, 2020 from https://github.com/acmsigsoft/open-science-policies/

Daniel Graziotin. 2018. How to disclose data for double-blind review and make it archived open data upon acceptance Retrieved July 12, 2020 from https://ineed.coffee/5205/how-to-disclose-data-for-double-blind-review-and-make-it-archived-open-data-upon-acceptance/.

Daniel Méndez, Daniel Graziotin, Stefan Wagner, and Heidi Seibold. 2019. Open science in software engineering. *arXiv*. https://arxiv.org/abs/1904.06499

GitHub. 2016. Making Your Code Citable. Retrieved July 12, 2020 from https://guides.github.com/activities/citable-code/. (How to automatically archive a GitHub repository to Zenodo)

Figshare. How to connect Figshare with your GitHub account. Retrieved July 12, 2020 from https://knowledge.figshare.com/articles/item/how-to-connect-figshare-with-your-github-account (How to automatically archive a GitHub repository to Figshare)

Sampling

An empirical study where some of many possible items are selected

Application

This standard applies to empirical research in which the researcher selects smaller groups of items to study (a *sample*) from a larger group of items of interest (the *population*) using a usually imperfect population list (the *sampling frame*). Common items in software engineering research include people (e.g. software developers), code artifacts (e.g. source code files) nd non-code artifacts (e.g. online discussions, user stories).

Specific Attributes

Section	Attribute
Essential	☐ explains the goal of sampling (e.g. aiming for representativeness, identifying exceptional cases) ☐ explains the sampling strategy, in particular the different filtering steps involved or the reasons for selecting certain objects ☐ explains why the sampling strategy is reasonable (not necessarily optimal) for the sampling goal ☐ explains the reasoning behind the selection of study objects (especially qualitative studies) ☐ reports the sample size
Essential only if representativeness is a goal	☐ states the theoretical population (what would the researcher like to generalize to?) ☐ presents a replicable, concise, algorithmic account of how other researchers could derive the same sample ☐ explicitly argues for representativeness (e.g. compares sample and population parameters, provides confidence interval and confidence level for sample size) ☐ explains how the sample could be biased along the sampling steps
Desirable	☐ reports the approximate or exact sizes of populations and sampling frames ☐ provides the sample, sampling frame, and sampling scripts as supplementary material (subject to the collected data containing sensitive or protected information). ☐ uses more sophisticated sampling strategies where appropriate, e.g.: • exploratory research: using purposive rather than convenience sampling for unit of analysis • case study: using purposive rather than convenience sampling for site selection • repository mining: using probability rather than convenience or purposive sampling (if a sampling frame is available) • online survey: using respondent-driven rather than snowball sampling • study with identifiable strata: using stratified random rather than simple random sampling • theory building: using theoretical rather than convenience sampling

Examples of Acceptable Deviations

- omitting a detailed account of the sampling strategy because it is explained in previous work using the same data set
- using a very simple sampling strategy in exceptional circumstances where expediency outweighs representativeness (e.g. research during a disaster)

Antipatterns

- making claims about a population, based on sample, without providing an argument for representativeness
- claiming that a sample is representative of a population because it was randomly selected from a sampling frame, without considering bias in the sampling frame
- conducting underpowered research; i.e.:
  - quantitative research with a sample size insufficient to detect effects of the expected size[21]
  - qualitative research with too little data for plausible saturation
- justifying the selection of items merely by stating that they come from a "real-world" context, without providing additional reasoning why the selected items are suitable for the study context

[21] Expected effect sizes should be plausible. For instance, expecting any single factor (e.g. programming language) to explain 50% of the variance in software project success is not plausible.

Invalid Criticisms

- complaining about lack of representativeness or low external validity in studies where representativeness is not a goal
- abstractly criticizing generalizability rather than pointing to best practices, e.g.:
  - invalid: ‘as most respondents work in app development, the results may not generalize to other settings’
  - valid: ‘the researchers should have sent participation reminders to mitigate response bias’
- for qualitative research, claiming that the sample size is too small without considering how the items were selected (e.g. theoretical sampling) or the authors’ argument for saturation.

Suggested Readings

Sebastian Baltes and Paul Ralph. 2020. Sampling in Software Engineering Research: A Critical Review and Guidelines. *arXiv*. https://arxiv.org/abs/2002.07764

William G. Cochran. 2007. Sampling techniques. *Wiley*.

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*. 285-311.

Barbara Kitchenham and Shari Lawrence Pfleeger. 2002. Principles of survey research: part 5: populations and samples. SIGSOFT Softw. Eng. Notes 27, 5 (September 2002), 17–20. DOI:10.1145/571681.571686

Gary T. Henry. 1990. *Practical sampling.* Sage 21.

Meiyappan Nagappan, Thomas Zimmermann, and Christian Bird. 2013. Diversity in software engineering research. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2013)*. Association for Computing Machinery, New York, NY, USA, 466–476. DOI:10.1145/2491411.2491415

Standards and Supplements under development

Artifact Evaluation

definition

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Case Survey

definition

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Discourse Analysis

definition

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting

Protocol Analysis

definition

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exploratory Data Science (Repository Mining)

definition

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Longitudinal Studies

Definition

Application

This standard applies to panel studies, cohort studies and other repeated measures studies …

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Simulation

Definition

Application

This standard applies to various levels of artificial environments: artificial (vs. "real-world") code, artificial bugs, artificial tests, artificial test failures, et. If the simulation involves direct observation of human participants, consider the **Protocol Study Standard**.

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Multi-Methodology (Supplement)

Definition.

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Replications (Supplement)

Definition.

Application

This standard applies to…

Specific Attributes

Importance	Attribute
Essential	☐
Desirable	☐
Extraordinary	☐

General Quality Criteria

Some criteria

Antipatterns

-

Examples of Acceptable Deviations

-

Invalid Criticisms

-

Notes

-

Suggested Readings

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Exemplars

Some references in ACM format: https://www.acm.org/publications/authors/reference-formatting.

Discussion

Background

Scholarly peer review is crucial to science: it not only determines what is published where but also who is hired, funded and promoted. Yet, virtually no one is happy with the current state of peer review. Every academic has peer review horror stories ranging from frustrating to emotionally devastating. While there is no silver bullet to solve every problem with peer review, this report presents a plan for mitigating some of the most serious problems.

Most journals and conferences provide general guidelines to authors and reviewers. Often, these guidelines ask reviewers to evaluate "soundness" but do not explain specifically what "methodologically sound" means for an experiment, case study, survey, literature review, etc. It is assumed that reviewers know what makes each methodology sound.

For each paper, each reviewer must therefore construct method-appropriate evaluation criteria. Reviewers may not have the necessary background to construct appropriate criteria. Even if the reviewers are familiar with the method, constructing a new evaluation system for each paper is just too much work. Instead, reviewers tend to proceed in a less systematic manner, simply highlighting problems that they notice. Even if reviewers did create more systematic scoring systems, reviewers, editors and authors may create diverse, incompatible systems with little relationship to each other, established methodological guidelines or community norms. This makes it impossible for a venue to communicate its expectations transparently to authors—there are no consistent expectations to communicate.

One solution is to separate discussing the meaning of "methodologically sound" from evaluating the soundness of a specific paper. That's what empirical standards do.

Standards-based Review

Publication venues may simply provide the standards and ask reviewers and editors to use them to evaluate manuscripts. This will help insofar as reviewers voluntarily stick to the standards. Most reviewers will probably use the standards most of the time because the standards make reviewing easier and reduce the cognitive load on the reviewer. However, some reviewers will continue going rogue—disregarding standards in favor of their own made up, non-consensus criteria. If authors meticulously craft studies and manuscripts to comply with standards that reviewers ignore, the review process will seem even more unjust and enraging. Therefore, venues should consider *standards-based review*.

In standards-based review, reviewers grade a paper using a structured form generated from relevant standards. The general process is as follows.

1) Each venue adopts a set of rules for mapping standards compliance into accept, revise or reject decisions. (A default mapping may be developed by the standards committee.)
2) During manuscript submission, the authors complete a submission checklist that determines which standards apply.
3) Someone (e.g. managing editor, submissions chair) completes the initial checks listed in the general standard and verifies that the correct standards have been selected.
4) A review form is automatically generated from the selected standards. Each attribute becomes one or a series of interconnected questions (see Figures 1 and 2 for examples). Space for free-form comments can be included at the discretion of the venue.
5) Two reviewers read the manuscript and answer the questions.

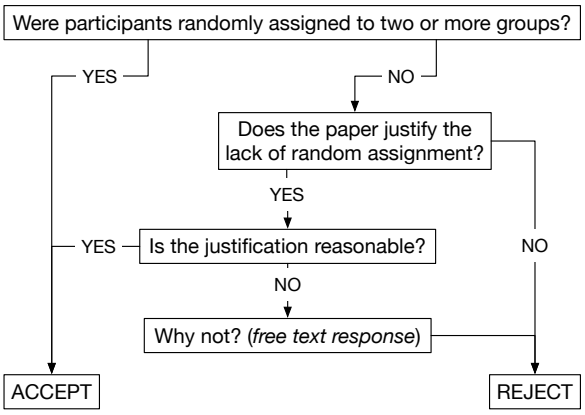

Figure 1: Random Assignment Decision Tree

6) (optional) Reviewers resolve specific disagreements (e.g. whether the paper checks assumptions of statistical tests) by discussion.
7) (optional) A third reviewer is recruited if the first two reviewers cannot reach a specified agreement threshold (does not have to be 100%; does not apply to all questions).
8) Reviewers do not make an overall recommendation (e.g. accept, major revision). Rather, the system compiles the reviewers' answers and tells the editor whether the paper meets the venue's rules for acceptance, revision or rejection. If a revision is needed, the system generates a to-do list for the authors based on the standards and the rules of the venue.
9) Rather than long, freeform reviews, authors receive the structured review results, which communicate acceptance, justify a rejection, or present a prioritized to-do list for a revision.

Challenges with Empirical Standards and Standards-based Review

Empirical standards are for directing, not replacing, expert judgment. The standards use words like "reasonable" and "convincing." They say things like "the limitations of the study are clearly acknowledged." A human being still has to judge whether things are reasonable or convincing. Expertise is still needed to determine whether any important limitations missing.

Standards-based review makes the review process more structured to improve consensus and prevent reviewers from inventing bogus criteria that the community has not agreed on (e.g. "I don't like this paper because the first author used a private email address"). However, the implementation of the standards should discourage a superficial, inflexible, box-ticking approach to review. One way to do this is using a decision-tree approach, at least for essential criteria. For example, Figure 1 illustrates a series of questions for random assignment (from the **Experiments with Human Participants Standard**). By specifically asking reviewers whether there is a reasonable justification for lack of random assignment, standards-based review can balance the competing needs for structure and flexibility.

A decision tree approach can help manage problems that are important but easily fixed in revisions. For example, Figure 2 shows possible pathways for evaluating a paper's limitations section. If the paper does not even attempt to explain the study's limitations, we reject. If the limitations section is adequate, we accept. If, however, the limitations section is inadequate, we ask the reviewers exactly what is incorrect or missing and invite a revision. This format encourages the reviewer to give actionable advice and transfers to the venue decisions regarding what problems lead to revision or rejection.

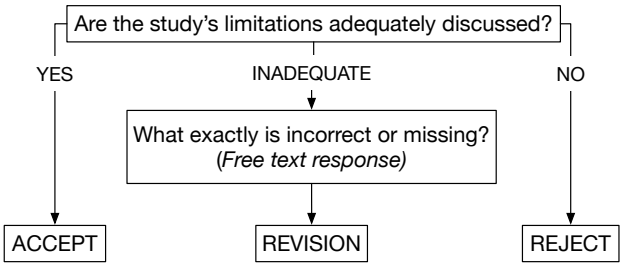

Figure 2: Limitations Decision Tree

Another possible problem is that reviewers may believe that a paper should be accepted despite violating the standards. This should not happen because the standards-based review system should ask whether each identified deviation from the standard is justified and reasonable. Papers that comply with relevant standards are accepted. When a paper does not comply, the reviewer is explicitly prompted to consider whether the deviation is makes sense. This process should discourage box-ticking.

The opposite problem is trickier. Reviewers may believe the paper should be rejected despite complying with the standards. Here, the desired workflow is to report the problem in the standards issue tracker (see below) so the standard can be updated, and, meanwhile, accept the paper. The paper should be accepted because rejecting papers over problems not mentioned in the standards is unfair to the authors and, if some critical flaw in one of the standards allows heaps of invalid research through, it will be quickly found and corrected. Allowing reviewers to override the standards at will would be rife for abuse; reviewers would carry on inventing bogus, non-consensus criteria. If the average reviewer could be so trusted, peer review would not be in its current dismal state.

Finally, the standards could create other unforeseen problems. To mitigate unknowns, venues could temporarily run traditional review alongside standards-based review, allowing authors to choose which kind they prefer, and comparatively evaluate the two systems. We can and should empirically study how the standards affect the review process.

Anticipated Benefits of Empirical Standards

An empirical standard is a model of a community's shared understanding of how a kind of study should be done. Providing these models to authors, reviewers and editors while implementing standards-based review should produce many benefits.

Benefits for Editors / Program Chairs

- improved review quality, thoroughness, consistency and readability
- reviewers cannot sit on the fence – the system always produces a clear recommendation
- more control over the review process and paper expectations; fewer rogue reviewers
- fewer angry emails from rejected authors
- easier to analyze inter-reviewer reliability and identify problem reviewers
- higher acceptance rates; faster publication times
- fewer low-quality submissions
- venues are perceived as being more fair, impartial and unbiased
- easier to recruit reviewers because reviews are less work

Benefits for Reviewers

- less overall reviewing work
    - reviewing is faster because there's no need to construct expectations, write an essay, copyedit the paper
    - less discussion among reviewers is needed; discussions can be more focused on disagreements between reviewers
    - third reviewer only needed if first two reviewers have insufficient agreement
    - fewer obviously deficient submissions because expectations are known
    - less re-reviewing manuscripts because more manuscripts are accepted straight away; revision to-dos are clearer, and most revisions can be checked by a single editor
    - all work that can be done by one person (e.g. checking keywords) is, and that one person is not a reviewer
    - less bouncing around of manuscripts between venues because many venues use the same standards and rejection rates are lower
- venue's expectations, for both reviewers and submissions, are clearer
- reviewers need less methodological expertise to review a paper effectively
- reduced extraneous cognitive load
- less arguing among reviewers: reviewers only have to determine whether a paper complies with a standard, not how, in principle, the study should have been conducted

Benefits for Researchers

- higher acceptance rates
- fewer major revisions
- faster publication times
- reviews are less emotionally charged, upsetting and demoralizing
- reviews are more actionable and consistent; no mixed messages
- reviewers cannot micromanage style
- reviewers cannot impose requirements inconsistent with established norms and guidelines
- venue's expectations are transparent to authors; reviews are more predictable
- no more superficial, "drive-by" reviews that do not engage with the paper's methodology and core claims
- standards provide convenient checklists for designing studies and polishing manuscripts
- standards mitigate bias against qualitative, exploratory and industry-focused research
- stops cross-paradigmatic criticism (e.g. "why doesn't this qualitative study have numbers?")
- standards are useful for training graduate students

Benefits to Society

- better, more rigorous, more ambitious studies
- less money and resources wasted because researchers forgot or ignored critical practices
- less HARKing and publication bias, because statistical significance is not a criterion
- less resistance to open reviewing / signed reviews (because reviews are less personal)
- less money and resources wasted on bad reviewing and re-reviewing
- fewer scientists and graduate students burning out due to destructive reviewing

Peer review is so frustrating not because reviewers find mistakes but because reviewers apply criteria that authors did not anticipate or do not agree with. Empirical standards increase transparency and separate debating research best practices from evaluating specific studies. Because the authors have the

standards, they can clearly see what reviewers are evaluating. They can make sure they have met each expectation and that it is obvious from their manuscripts. Reviewers can in turn check whether the manuscript meets each expectation. Authors cannot sidestep essential practices and reviewers cannot invent unexpected criteria. Furthermore, standards give reviewers who are less familiar with a methodology a better chance of both appreciating what was done well and noticing fundamental problems.

Myth: peer review is upsetting because reviewers find authors' mistakes
Truth: peer review is upsetting because authors and reviewers disagree on appropriate research practices

Moreover, extracting the process of creating the standard from the process of reviewing a specific paper facilitates a broad debate about how studies should be performed and presented, which encourages consensus. Meanwhile, a standards-based review process prevents reviewers from applying rules or expectations that authors cannot anticipate or that do not reflect methodological consensus. This will make the review process fairer, more consistent, more transparent and less emotionally damaging. Since authors know what reviewers are looking for in advance, acceptance rates should rise.

How we Created the Standards

In 2018, ACM SIGSOFT solicited volunteers, using the SEWORLD mailing list and social media, for a series of special initiatives. Inclusion was not limited to SIGSOFT members. Paul Ralph (*Dalhousie University*) and Romain Robbes (*Free University of Bozen-Bolzano*) were asked to co-chair the *Paper and Peer Review Quality* initiative. Paul focused on developing empirical standards while Romain investigated possible improvements to peer review processes. (After some research, it became clear that many groups are working on improving the peer review process, so the Initiative is now chiefly focused on empirical standards).

After the initiative leaders were announced at the ICSE town hall in May 2019, Paul contacted the volunteers and shared plans for drafting empirical standards. Volunteers:

- reviewed the plan for drafting the standards;
- discussed which methods should have a standard;
- created a template illustrating the sections each standard should have;
- suggested additional researchers who might want to volunteer; and
- nominated subject matter experts to draft each standard.

Over forty people eventually joined the *Paper and Peer Review Quality Task Force.*

Each standard was assigned to between one and three subject matter experts who produced a draft, based on existing published guidance where possible. Drafts included some or all of the following items.

- Name of method (e.g. Experiment, Questionnaire Survey, Case Study)
- Definition of the method
- Application (i.e. where the standard does and does not apply)

- Specific Attributes (i.e. the properties of a study that make it publishable), organized into categories such as *essential*, *desirable* and *extraordinary* (defined below)
- General Quality Criteria (e.g. internal validity, construct validity, credibility, resonance)
- Examples of Reasonable Deviations (to emphasize that standards are flexible)
- Antipatterns (i.e. common problems to watch out for)
- Invalid criticisms (i.e. things reviewers should not say)
- Suggested readings (including longer methodological guidance on which standards are based)
- Exemplars (good examples of the method authors should emulate)
- Notes (important clarifications that did not fit elsewhere)

The drafts were then edited for consistency, condensed, and compiled into this document. Next, this document, including the standard drafts, was circulated to the task force for feedback, and updated accordingly. Then we created a GitHub repository to house this report and through which issues can be reported. Feedback will be solicited from the community via social media and the SEWORLD listserv.

Currently, we have standards for research methods that are used in software engineering, and for which we could find subject matter experts willing to develop a standard. More standards are planned, and still more may be needed.

Principles

The standards are guided by the following principles.

- Manuscripts should be evaluated on their own terms, vis-à-vis their context and goals.
- The scientific community (not individual reviewers) should determine expectations and norms.
- Expectations of journals and conferences should *reflect* community consensus.
- Researchers should follow standards *where they make sense, and justify reasonable deviations*.
- Reviewers should focus on the manuscript's methodological soundness.
- Reviewers should **not** micromanage style or copyedit.
- Reviewers should evaluate adherence to best practices (e.g. using validated scales), **not** abstract criteria (e.g. construct validity)
- The standards *must not be biased* against specific research methodologies or areas of interest.

Interpreting and Applying the Standards

The empirical standards described here are intended for evaluating complete empirical studies. For now, at least, pilot studies and non-empirical scholarship are not included. Venues should consider the possibility of substantially scaling back peer review for pilot studies, posters, short papers, workshop papers and position papers to control reviewing loads.

Advantageous attributes of manuscripts and studies are divided into categories. **Essential attributes** are necessary conditions for publishing the work. Without a compelling justification, a study that does not meet one or more essential attributes should not be published in any peer-reviewed venue. **Desirable attributes** are recommended but not always necessary or applicable. Some desirable attributes may be mutually exclusive. For publication in a prestigious venue, a study should exhibit some, but not all, desirable attributes. The presence of desirable attributes is what distinguishes high-quality research. **Extraordinary attributes** are not expected and indicate award-quality research. One reason the standards describe extraordinary attributes is to emphasize that these attributes should not be expected or taken for granted.

The **General Standard** applies to all empirical research. When no more specific standard is available, reviewers can fall back on the general standard.

Most manuscripts, however, should be evaluated using multiple standards. For instance, a multimethodological study combining a systematic literature review with a questionnaire survey would be evaluated against the **general standard**, the **systematic review standard**, the **questionnaire survey standard** and the **multi-methodology supplement**. We envision a system that automatically combines the relevant standards during the submission process, and automatically assembles them into a reviewing form.

The absence of an attribute indicates that it should not be considered. For example, **The General Standard** does not say that papers should be free from arbitrary decisions because most research involves some arbitrary decisions and their presence does not invalidate findings.

We have also included several **Supplements** for cross-cutting concerns like information visualization and sampling. These supplements help to reduce duplication between the standards, and give more information about complex issues.

Evolving and Governing the Standards

The empirical standards are living documents, which should be continuously revised to reflect evolving consensus around research best practices. Issues can be reported and changes can be requested via GitHub.[22]

Each standard will require one or more *maintainers*. Each maintainer will be responsible for reviewing issues and change requests filed against their standard, and will have the authority to make, accept or reject changes. Crucially, maintainers must be experts *in that kind of research* to avoid misapplying norms from one kind of study to a fundamentally different kind. To become a maintainer, therefore, a person must:

- Have a PhD or equivalent terminal degree in software engineering, computer science or a related discipline.
- Have published one or more studies related to the standard (e.g. for the systematic reviews standard, a maintainer must have published a paper that reports a systematic review, analyzes a sample of systematic reviews, or provides guidelines for performing systematic reviews) in a prestigious software engineering journal or conference in the past six years.
- Be approved by the *steering committee*.

The *steering committee* will consist of:

- The director of the empirical standards project—currently Paul Ralph (Dalhousie University)
- The editors-in-chief (or their designated representatives) of all of the DBLP-indexed journals that adopt the standards
- A designated representative of the steering committee of each DBLP-indexed conference that adopts the standards

[22] https://github.com/acmsigsoft/EmpiricalStandards

- Two, elected, early-career members-at-large

The steering committee will determine how it appoints maintainers and can institute policies to exclude illegitimate or predatory venues or include someone who, in hindsight, should have been included. Once the initial steering committee is formed, it will be responsible for elaborating and formalizing the governance model, defining "early-career" and determining how early-career members should be elected.

Justice, Equality, Diversity and Inclusion

The current system of peer review is demonstrably biased in many ways.[23, 24] One purpose of developing these empirical standards is to make peer review less biased. However, the more a person has suffered from these biases, the less likely they would be involved in creating and implementing these standards. To mitigate bias, we took the following steps:

- anyone could volunteer for the task force
- all volunteers were encouraged to work on one or more standards
- we specifically tried to recruit women, people of color and members of other underrepresented groups to be involved with the standards
- more than 25 subject matter experts were involved in drafting the standards; more than 40 researchers reviewed the initial drafts
- different methods have different standards, reducing bias against specific approaches
- we tried to write make general standard, especially, philosophy- and method-agnostic
- the standards will be released for extensive public comment before being implemented
- we will use GitHub to make all feedback and changes transparent
- we added a provision for members at large to get a junior perspective on the steering committee
- anyone will be able to report an issue in or suggest an improvement for a standard

These steps are likely insufficient to eliminate bias in the standards but are the best we could come up with. Further suggestions are welcome at any time.

Disclaimers

Nothing in these standards should be interpreted as privileging any methodology (e.g. controlled experiment vs. ethnomethodology), epistemological position (e.g. positivism vs. interpretivism), analysis approach (i.e. frequentist vs. Bayesian statistics), or data type (e.g. quantitative vs. qualitative). Each study should be reviewed on its own terms, which is why so many different standards are needed.

Acknowledgements

The editor would like to acknowledge the assistance of ACM SIGSOFT, and president Thomas Zimmerman in particular, for kickstarting this initiative.

[23] Giangiacomo Bravo, Mike Farjam, Francisco Grimaldo Moreno, Aliaksandr Birukou, and Flaminio Squazzoni. 2018. Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics* 12, 1: 101-112.
[24] Paul Ralph. 2016. Practical suggestions for improving scholarly peer review quality and reducing cycle times. *Communications of the Association for Information Systems* 38, 1: article 13.

Glossary

Artifact: an artificial object (e.g. computer program, document, method, model, practice, technique, template)

Chain-of-evidence: (in qualitative research) a mapping of raw data (e.g. quotations) to theoretical concepts (e.g. themes, categories), typically with one or more intermediate steps (e.g. codes, labels, subcategories), sometimes presented as a table.

Credibility: the extent to which conclusions are supported by rich, multivocal evidence.

Construct validity: Do the measures support the research objective? The questionnaire items (questions) and related response scales should accurately represent the research aims.

External validity: Can the conclusions be generalized to the target population? The characteristics and size of the sample should represent the extended population.

Generalizability: See **External validity**.

Internal validity: Are the relationships between the investigated factors examined? In survey research, it is difficult to control the conditions in which the factors are studied and to account for potential confounding factors. Low internal validity is expected.

Multivocal: The property of being based on—and recognizing differences between—people with different opinions and backgrounds (including gender, culture, education, and class).

Objectivity: Are the results free from the bias of the researchers? This can be achieved through standardization of the procedures for data collection, analysis, and interpretation.

Recoverability: A study is recoverable when readers can understand how the work was done and why it was done that way. All research should be recoverable.

Reflexivity: the extent to which authors reflect on their potential biases and interactions with the team, organization or community, especially possible negative impacts on some participants or stakeholders.

Reliability: Can the reviewer arrive at similar results if applying the described procedures? Through reproducing the data analysis, we expect the same results and a similar interpretation of the evidence.

Replicability: A study is replicable, when the data collection and analysis (on the new data) can be repeated by an independent researcher. Positivist research should be replicable; interpretivists and postmodernists reject the notion that social science is replicable. Qualitative research is typically not replicable.

Reproducibility: A study is reproducible when an independent researcher can precisely recreate the results using the original study's data and source code. Interpretivists and postmodernists reject the notion that social science is reproducible, and qualitative research is typically not reproducible. Much positivist research is not reproducible because it is impractical or unethical to publish the dataset.

Resonance: the extent to which a study's conclusions make sense to (i.e. resonate with) participants

Rigor: the extent to which theory, data collection, and data analysis are sufficient, appropriate and not oversimplified.

Site: the conceptual space within which a study's data collection occurs. In qualitative and especially case study research, the concept of site is not limited to physical location, but rather defines the boundaries within which the research takes place.

Theoretical sampling: choosing which data to collect based on the emerging theory, concepts or categories; typically used in qualitative research, especially Grounded Theory.

Transferability: the extent to which a study's results could plausibly apply to other sites, people or circumstances.

Usefulness: the extent to which a study provides actionable recommendations to researchers, practitioners **OR** educators.

List of Contributors

Editor

Paul Ralph, *PhD (British Columbia), B.Sc. / B.Comm (Memorial),* is an award-winning scientist, author, consultant and Professor of Software Engineering at Dalhousie University in Halifax, Canada. Paul co-chairs the ACM SIGSOFT Paper and Peer Review Quality Initiative and has written extensively on research methodology for software engineering.

Contributors

Sebastian Baltes, University of Adelaide, Australia

Domenico Bianculli, University of Luxemburg, Luxemburg

Yvonne Dittrich, IT University of Copenhagen, Denmark

Neil Ernst, University of Victoria, Canada

Michael Felderer, University of Innsbruck, Austria

Robert Feldt, Chalmers University of Technology, Sweden

Antonio Filieri, Imperial College London, UK

Carlo Alberto Furia, USI Lugano, Switzerland

Daniel Graziotin, University of Stuttgart

Pinjia He, ETH Zurich, Switzerland

Rashina Hoda, Monash University, Australia

Natalia Juristo, Technical University of Madrid, Spain

Barbara Kitchenham, Keele University, UK

Romain Robbes, Free University of Bozen-Bolzano, Italy

Daniel Mendez, Blekinge Institute of Technology, Sweden, and fortiss GmbH, Germany

Jefferson Molleri, Simula Metropolitan Centre for Digital Engineering, Norway

Janet Siegmund, Chemnitz University of Technology, Germany

Diomidis Spinellis, Athens University of Economics and Business, Greece

Miroslaw Staron, University of Gothenburg, Sweden

Klaas Stol, University College Cork, Ireland

Damian Tamburri, Eindhoven University of Technology, Netherlands

Marco Torchiano, Polytechnic University of Turin, Italy

Christoph Treude, University of Adelaide, Australia

Burak Turhan, Monash University, Australia

Sira Vegas, Technical University of Madrid, Spain