Background. Software Engineering (SE) research, like other applied disciplines, intends to provide trustful evidence to practice. To ensure trustful evidence, a rigorous research process based on sound research methodologies is required. Further, to be practically relevant, researchers rely on identifying original research problems that are of interest to industry; and the research must fulfill various quality standards that form the basis for the evaluation of empirical research in SE. A dialogue and shared view of quality standards for research practice is still to be achieved within the research community.

Objectives. The main objective of this thesis is to foster dialogue and capture different views of SE researchers on method level (e.g., through the identification and reasoning on the importance of quality characteristics for experiments, surveys and case studies) as well as general quality standards for Empirical Software Engineering (ESE). Given the views of research quality, a second objective is to understand how to operationalize, i.e. build and validate instruments to assess research quality.

Method. The thesis makes use of a mixed method approach of both qualitative and quantitative nature. The research methods used were case studies, surveys, and focus groups. A range of data collection methods has been employed, such as literature review, questionnaires, and semi-structured workshops. To analyze the data, we utilized content and thematic analysis, descriptive and inferential statistics.

Results. We draw two distinct views of research quality. Through a top-down approach, we assessed and evolved a conceptual model of research quality within the ESE research community. Through a bottom-up approach, we built a checklist instrument for assess survey-based research grounded on supporting literature and evaluated ours and others’ checklists in research practice and research education contexts.

Conclusion. The quality standards we identified and operationalized support and extend the current understanding of research quality for SE research. This is a preliminary, but still vital, step towards a shared understanding and view of research quality for ESE research. Further steps are needed to gain a shared understanding of research quality within the community.
Views of Research Quality
in Empirical Software Engineering

Jefferson Seide Molléri
Views of Research Quality in Empirical Software Engineering

Jefferson Seide Molléri

Doctoral Dissertation in Software Engineering

Department of Software Engineering
Blekinge Institute of Technology
SWEDEN
### CONTENTS

#### 3 CERSE - Catalog for empirical research in software engineering: a systematic mapping study
- 3.1 Introduction ............................................ 69
- 3.2 Related Work ........................................... 71
- 3.3 Method .................................................. 73
- 3.4 Results .................................................. 84
- 3.5 Discussion .............................................. 100
- 3.6 Conclusions ............................................. 107
- 3.7 References ............................................. 110
- 3.8 Appendix ............................................... 112

#### 4 Aligning the Views of Research Quality in Empirical Software Engineering
- 4.1 Introduction ............................................ 175
- 4.2 Background & Related Work .......................... 177
- 4.3 Internal Case Study .................................... 180
- 4.4 Focus Group ............................................ 189
- 4.5 Discussion .............................................. 203
- 4.6 Conclusion .............................................. 212
- 4.7 References ............................................. 214
- 4.8 Appendix ............................................... 218

#### 5 Teaching students critical appraisal of scientific literature using checklists
- 5.1 Introduction ............................................ 221
- 5.2 Background ............................................. 223
- 5.3 Research method ....................................... 224
- 5.4 Results .................................................. 230
- 5.5 Discussion .............................................. 237
- 5.6 Conclusions ............................................. 241
- 5.7 References ............................................. 241
- 5.8 Appendix ............................................... 248

#### 6 An Empirically Evaluated Checklist for Surveys in Software Engineering
- 6.1 Introduction ............................................ 251
- 6.2 Background & Related Work .......................... 253
CONTENTS

6.3  Step 1. Construction of the checklist .......................... 256
6.4  Step 2. Evaluation of the checklist ............................ 271
6.5  Discussion ......................................................... 281
6.6  Conclusions ....................................................... 285
6.7  References ......................................................... 286
6.8  Appendix ............................................................ 293
Abstract

Background. Software Engineering (SE) research, like other applied disciplines, intends to provide trustful evidence to practice. To ensure trustful evidence, a rigorous research process based on sound research methodologies is required. Further, to be practically relevant, researchers rely on identifying original research problems that are of interest to industry; and the research must fulfill various quality standards that form the basis for the evaluation of the empirical research in SE. A dialogue and shared view of quality standards for research practice is still to be achieved within the research community.

Objectives. The main objective of this thesis is to foster dialogue and capture different views of SE researchers on method level (e.g., through the identification and reasoning on the importance of quality characteristics for experiments, surveys and case studies) as well as general quality standards for Empirical Software Engineering (ESE). Given the views of research quality, a second objective is to understand how to operationalize, i.e. build and validate instruments to assess research quality.

Method. The thesis makes use of a mixed method approach of both qualitative and quantitative nature. The research methods used were case studies, surveys, and focus groups. A range of data collection methods has been employed, such as literature review, questionnaires, and semi-structured workshops. To analyze the data, we utilized content and thematic analysis, descriptive and inferential statistics.

Results. We draw two distinct views of research quality. Through a top-down approach, we assessed and evolved a conceptual model of research quality within the ESE research community. Through a bottom-up approach, we built a checklist instrument for assessing survey-based research grounded on supporting literature and evaluated ours and others’ checklists in research practice and research education contexts.
Conclusion. The quality standards we identified and operationalized support and extend the current understanding of research quality for SE research. This is a preliminary, but still vital, step towards a shared understanding and view of research quality for ESE research. Further steps are needed to gain a shared understanding of research quality within the community.

Keywords: Research Quality, Quality Standards, Empirical Software Engineering, Research Methodology
Acknowledgements

Firstly, I would like to express my appreciation to my supervisors Professor Kai Petersen and Professor Emilia Mendes for their equally valuable guidance, thoughtful advice, and for accepting nothing less than excellence from me. Kai has provided me continuous assistance, collaboration and ‘Tolkienian’ friendship. Emilia has taught me a vision of integrity, a philosophy that has greatly influenced my approach to my research. They both have inspired me, and I hope to continue improving and learning from them.

I am also thankful to my colleagues at SERL-Sweden for their feedback, cooperation and of course friendship. In particular, I would like to thank Nauman bin Ali and Javier Gonzalez-Huerta for insightful discussions, and Bogdan Marculescu and Raquel Ouriques for the lighthearted moments. I would also like to express my gratitude to Monica Nilsson, Professor Claes Wohlin and Professor Jürgen Börstler for vital advice in overcoming numerous obstacles.

I will always be grateful that my first role models - my parents José Renato and Gertrudes - provided my siblings and me an environment to be truthful and honest. And also to my siblings and closest friends, for their invaluable support and encouragement. To all the friends in Brazil, who offer creativity and conflict resolution exercises, and to promote my character growth. Last but not least, I would like to extend my deepest gratitude to Mei, for her kind-heartedness and understanding when I need it the most.

This work is supported by the Science Without Borders program, funded by CNPq (National Council for Scientific and Technological Development - Brazil).
Overview of Papers

Papers in this Thesis


**Contribution Statement**

Jefferson Seide Molléri is the first author for all the papers that compose this Thesis. As lead author, Jefferson was responsible for designing and conducting the studies, collecting and analyzing data, and most reporting activities. In
addition, he is the sole author of Chapter 1, the Introduction. The detailed authors’ contribution to the chapters are described next:

**Chapter 2:** Jefferson Seide Molléri and Kai Petersen conceived the presented idea. Jefferson Seide Molléri collected the data and performed the computations. Kai Petersen and Emilia Mendes verified the analytical methods and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

**Chapter 3:** All authors contributed to conceiving the idea and planning the research. Jefferson Seide Molléri conducted the search process and collected data from the candidate papers. All the authors took part in the study selection process. Kai Petersen and Emilia Mendes also verified the data extraction and synthesis of included studies. Finally, all authors discussed the findings and contributed to the final manuscript.

**Chapter 4:** Jefferson Seide Molléri, Emilia Mendes and Kai Petersen conceived the idea and planning the overall research. Jefferson Seide Molléri designed and carried out the case study. Jefferson Seide Molléri and Michael Felderer designed and conducted the focus group. Kai Petersen and Emilia Mendes verified the data collection and synthesis. Jefferson Seide Molléri analyzed the data from both studies. All authors discussed the results and contributed to the final manuscript.

**Chapter 5:** Nauman bin Ali and Kai Petersen conceived the original idea and Jefferson Seide Molléri planned and led the research. Jefferson Seide Molléri and Nasir Mehmood Minhas designed and conducted the evaluation task with the students. Jefferson Seide Molléri, Nasir Mehmood Minhas, and Panagiota Chatzipetrou analyzed the data and interpreted the results. All authors contributed to the final manuscript.

**Chapter 6:** All authors contributed to conceiving the idea and planning the research. Jefferson Seide Molléri carried out the construction and evaluation of the checklist. Kai Petersen and Emilia Mendes verified the data extraction and synthesis of guidelines. All authors discussed the results and contributed to the final manuscript.

**Related Papers not in this Thesis**

**Paper 1:** Jefferson Seide Molléri, Fabiane Benitti. “SESRA: a web-based automated tool to support the systematic literature review process”. In *Proceedings*


Chapter 1

Introduction

1.1 Overview

Empirical research contributes towards knowledge by means of a creative and systematic investigation of theories and hypothesis. It comprises a structured process to collect, analyze and interpret evidence. This structured, evidence-based approach can take many forms, such as experiment, case study, and survey [23, 38]. To ensure the validity and trustfulness of the resulting evidence, the research process should be carried out with high methodological rigor [5, 8].

Applied sciences, such as Software Engineering (SE), value evidence-based approach to generate knowledge and to devise novel applications [2, 8]. This denotes a commitment towards practical relevance, i.e. the potential to transfer research results to practice in order to improve the software development process and related activities. Relevant research is also likely to impact academia and guide the research community towards the phenomena worthy to investigate [2].

In the context of software engineering, novel applications may refer to new solutions to software engineering problems. Through empirical evaluations the solutions become more relevant, as benefits and limitations may be demonstrated. Thus, empirical evaluations also play an important role in the transfer of research results to practice in order to improve the software development process and related activities (see e.g. [14]).

Methodological rigor and practical relevance are properties (i.e. qualities) of research excellence. Research that aims at excellence by maximizing the desirable qualities is regarded as right or proper. This notion of excellence
is complemented by secondary properties such as perceived interest, originality, and credibility of research [18, 27]. However, different stakeholders (e.g., research collaborators and research institutions) are likely to regard research qualities in different degrees, according to personal opinions, role perspectives, and expectations [28].

Rather than reward excellence in research with regard to its primary qualities, studies are often evaluated by citation indices that measure academic impact based on publications. Today these indices play an essential role to rate institutions, researchers, and their academic performance; however, the use of such indices faces several criticisms, in particular, that the citation count slows scientific progress [10, 29].

In an ideal scenario, the quality of research is characterized by both methodological rigor and attested practical relevance. If this is not the case, then such quality misalignment should be made explicit, so to encourage the research community to adequately acknowledge good quality in research, and regarding studies that show excellence in both characteristics.

Reaching such ideal scenario poses challenges for empirical research, such as: (i) how to measure the research quality alignment between different stakeholders; (ii) what are the effects of (mis)alignment with regard to the research process and produced results; and (iii) how to improve such quality’s alignment.

With this thesis, we propose to investigate existing gaps and bonds with regarding different views of research quality in empirical software engineering (ESE). In particular, we: investigate the relationship between research quality and academic impact, and determine the alignment of experts with regard to desired research qualities. Furthermore, we intend foster improvements in the research quality by providing: a catalog of guidelines and supporting instruments to research practice, and empirically evaluated instruments to assess the methodological quality of SE research.

The remainder of this chapter is structured as follows. Background and related work are presented in Section 1.2, followed by the research contribution and methodology in Section 1.3. Further, Section 1.4 details the studies included as part of this thesis. A discussion about the contributions of this Thesis for research practice are provided in Section 1.5, followed by Conclusions and ideas for further work in Section 1.6.
1.2 Background and related work

Research practice is conducted in accordance with methodological guidelines that incorporate the principles of excellence [27]. Such guidelines intend to direct the conformance between the research process and research outcomes with desired quality standards. This section provides an overview of the different perspectives of research quality in the SE context.

1.2.1 Research Quality

Primary qualities of Software Engineering research are often associated with a rigorous research process and a relevant application to practice [17,18,33]. These two dimensions comprise the core of research quality in the field and guide the assessment of empirical research [18]. In addition to these two dimensions, research has also been evaluated through a series of secondary qualities, such as excellence in science communication and conformance to ethical standards [27].

Methodological rigor addresses the validity of research, ensuring that the research process is reasonable and correct according to the proposed methodology [9,18]. Methodological quality standards are focused on the research practice, covering design and execution of research studies. It is therefore centered on the role of the researcher, and the actions that the researcher is responsible for in order to achieve excellence [28].

The impact of research is not restricted to the academic setting. When correctly conducted, research results are confirmed in a context, and therefore applicable in that particular setting [25]. Thus, the interest and needs of the target audience are decisive for the practical relevance and originality of a study. To ensure that the potential to impact the field is not wasted, high-quality research is aligned with the trends of industrial practice [2,12].

To achieve such an impact, the community expects researchers to transfer the knowledge produced by their studies to practice. Excellence in communicating research depends on a relationship between researcher and publisher but also community and editors [33]. On one side, a researcher should write a report in a style that the audience would understand and enjoy reading [2]. As many journals have researchers as their target audience (who are also reviewers) the communication style is not geared towards practitioners. On another side, the publisher is responsible for structuring a body of knowledge, and to make it accessible to the target audience [27,28].

Finally, research is not confined to the scientific environment alone. Ultimately, research results should be of interest to the overall community, and
therefore one expects high-quality research to advance human knowledge, culture, and society [28]. Because of that, research practice and research outcomes should comply with standards for ethics and conformance to social, economic and environmental regulations [35], and to take into account concerns and values of both the target audience and research stakeholders [28].

1.2.2 Models of Research Quality

Attempts to define research quality are numerous, proposed as models to describe quality of research in diverse research fields, and via multidisciplinary initiatives.

A set of Criteria for Assessing the Trustworthiness of Naturalistic Inquiries is described by Guba [15]. The main four aspects (i.e. truth value, applicability, consistency and neutrality) are presented side by side with their scientific and naturalist counterparts. The scientific criteria are more closely associated to a realist or positivist paradigm of research, while as the naturalist view tends towards the relativist or qualitative research. These traditional research criteria are further used benchmark for models of research quality and validity frameworks [3,9,31].

As an example, the Encyclopedia of Qualitative Research [13] provides an extensive list of qualities associated with qualitative research. The compendium provides summaries about quality concepts such as rigor and relevance and related topics. As an example, rigor in qualitative research is associated with credibility, dependability, reflexivity, transparency, and trustworthiness. Very often, qualitative aspects of research are described in comparison to quantitative research (e.g. transferability and generalizability). The encyclopedia does not provide a conceptual model that synthesizes its comprehensive set of dimensions.

Benbasat and Zmud [2] detail a list of Dimensions of Relevance for Information Systems (IS) research. The list comprises only four key dimensions: interesting, applicable, current and accessible. Although focused upon the relevance aspect alone, this set provides a solid ground for discussions on research excellence and recommendations to achieve it. Moreover, the relevance dimensions were later aggregated to other dimensions of quality, leading to a more comprehensive model of research quality [26].

Mårtensson et al. [26,27] proposed and evolved a Multidisciplinary Model of Research Quality. It aggregates the knowledge from diverse fields, such as medicine, information systems, social sciences and business to draw up a conceptual hierarchy of quality dimensions. Initially, the model comprised three
broad quality dimensions (i.e., credible, contributory and communicable), later also including ‘conforming to regulations’ and ‘standards’.

This multidisciplinary model attempts to describe research quality combining different perspectives, such as quantitative and qualitative research approaches, and both social and empirical research. In order to be operationalized though, the authors encourage the evaluation of the conceptual model across specific domains. The evaluation is also intended to foster a discussion about measures to be collected, and prioritization of dimensions according to the aims of operationalization.

The Research Quality Plus Assessment Framework – RQ+ [28] is a comprehensive approach to characterize and assess research quality. It comprises a set of core quality dimensions (e.g., integrity, legitimacy, importance, and positioning for use) and related sub-dimensions. Its intention is to help researchers and institutions, through a systematic process, to synthesize a contextual instrument to assess research quality and performance.

RQ+ also describes a conceptual model of research practice and its evolution from execution to the application of research outputs. Initially, the research is closely related to researchers and other contributors. Aspects of methodological quality, such as credibility and potential relevance are more easily controlled here. On a further step, the originality of research is evaluated through its sphere of influence. Finally, impact to practice, as well as to sociocultural, socioeconomic and environmental determinants are perceived.

The five models we describe herein present similarities and differences with regarding the dimensions of research quality they comprise, as illustrated in Table 1.1. None of the quality models comprises a complete set of the dimensions. Most of the models address aspects related to the traditional scientific research criteria [9, 15]. One exception is the model focused only on dimensions of relevance [2].

The models cover dimensions of research quality with different levels of detail. As an example, the dimension accessible proposed by Benbasat and Zmud [2] is represented by three more specific dimensions in Mårtensson et al. [27]. In addition to that, some aspects of research quality (e.g. timeless and actionability) are mentioned by one of the models only.
Table 1.1: Summary of the models of research quality and related dimensions. Similar dimensions are presented in the same table row.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth Value</td>
<td></td>
<td>Credible</td>
<td>Integrity</td>
<td></td>
</tr>
<tr>
<td>Internal validity</td>
<td>Rigor in qualitative research, validity</td>
<td></td>
<td>Rigorous, internally valid, reliable</td>
<td></td>
</tr>
<tr>
<td>Credibility</td>
<td>Credibility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contextuality</td>
<td></td>
<td>Contextual, coherent</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contributory</td>
<td>Importance</td>
<td></td>
</tr>
<tr>
<td>Applicability</td>
<td></td>
<td>Relevance</td>
<td>Relevant</td>
<td>Relevance</td>
</tr>
<tr>
<td>External validity</td>
<td>Validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalizability</td>
<td>Generalizability</td>
<td></td>
<td>Generalizable</td>
<td></td>
</tr>
<tr>
<td>Transferability</td>
<td>Transferability</td>
<td></td>
<td>Generalizable</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interesting</td>
<td>Relevant research idea</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Applicable</td>
<td>Applicable result</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Current</td>
<td>Current idea</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Original, original idea, original procedure, original result</td>
<td></td>
</tr>
<tr>
<td>Authenticity</td>
<td></td>
<td></td>
<td>Originality</td>
<td></td>
</tr>
<tr>
<td>Consistency</td>
<td></td>
<td></td>
<td>Reliable, consistent</td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependability</td>
<td>Dependability, trustworthiness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutrality</td>
<td>Reflexivity</td>
<td></td>
<td>Transparent</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
### 1.2 Background and related work

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectivity</td>
<td>Objectivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmability</td>
<td>Confirmability</td>
<td>Communicable</td>
<td>Positioning for Use</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consumable, structured, understandable</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accessible</td>
<td>Readable, accessible, searchable</td>
<td>Knowledge accessibility and sharing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Timeliness and actionability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrity</td>
<td></td>
<td>Conforming</td>
<td>Legitimacy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aligned with regulations, ethical, morally justifiable, open</td>
<td>Addressing potentially negative consequences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equal opportunities</td>
<td>Gender-responsiveness, inclusiveness, engagement with local knowledge</td>
<td></td>
</tr>
<tr>
<td>Accountability</td>
<td></td>
<td>Sustainable</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 1.2.3 Assessing the Quality of Software Engineering Research

Many different instruments have been proposed to assess SE studies, with regard to their methodological quality and the quality of resulting evidence. Instruments focused on the methodological quality assess whether the process is strictly conducted, while the instruments focused on the evidence assess the quality of reporting and findings.
Instruments Assessing the Quality of Evidence

In their article “Strength of evidence in systematic reviews in software engineering,” Dybå, and Dingsøyr [8] presented an overview of some of the most influential instruments for assessing the quality of evidence across other fields, especially healthcare. The authors also discuss how to apply such instruments in the context of SE research for grading the overall strength of a body of evidence.

Dybå and Dingsøyr [7] also propose an instrument to assess the quality of primary studies during an Systematic Literature Review (SLR). The instrument is composed of eleven criteria, from which the three first ones are based on the study report and define a threshold, i.e. studies that do not comply with these minimum set of criteria are therefore excluded. The remaining criteria are based on the Critical Appraisal Skills Programme (CASP\(^1\)) and address methodological rigor, credibility of the findings, and relevance to practice.

This instrument is later combined to another [21] in order to produce a more complete one. The combined instrument was then used to evaluate whether researchers can assess fairly the quality of empirical studies by means of a quality evaluation checklist [20]. Results from the evaluation point out a poor degree of reliability of individual assessments, thus suggesting that multiple reviewers are required even with the support of an assessment checklist.

Instruments Assessing Methodological Quality

Runeson and Höst [34] proposed guidelines for conducting and reporting case study research in SE, alongside a set of checklists for design, data collection, reporting and reviewing case studies. The reviewing checklist assesses whether case studies comply with twelve practices recommended by the guidelines. Similar to that, Wohlin et al. [38] introduce in the guidelines for experimentation a checklist with 23 questions for reviewing and assessing experimental studies. Both publications explicitly define the checklists, highlighting the need for developing and providing such assessment instruments in addition to the description of the guidelines.

Ivarsson and Gorschek [17] created an assessment rubric focused on the primary qualities of software engineering research, namely rigor and relevance. The methodological rigor reflects how the research publication is conducted and reported, whereas relevance refers to the potential impact of the research on practice. The instrument is not intended to assess the research produced with a

\(^1\)https://casp-uk.net/.
specific research method but gives more importance to research that investigate real situations, e.g., action study, case study and interviews with practitioners.

Instruments assessing the methodological quality are often tailored with respect to a research method, such as experiment or case study research [36,37]. A common problem is that using a single instrument to assess the quality of empirical studies that employ different research methods is difficult. The need for an instrument that can assess studies across different research methods has been highlighted [36,37]. Such an instrument intends to provide a common ground for evaluation and comparison of research quality regardless of the methodology. Thus, the authors constructed [36] and evaluated [4,37] a unified checklist based on other existing instruments for assessing experiments and case studies.

Kitchenham and Brereton [19] also encourage the use of a common set of criteria to compare the methodological quality of papers employing different research methods. However, this proved problematic as each researcher can classify the empirical study differently (e.g., a case study based on an opinion survey, could be classified in either type), and some questions are inappropriate for certain types of study.

According to the evaluation conducted by Kitchenham and Brereton [19], there are potential limitations when using a common checklist to assess the variety of different empirical methods. As an example, small studies could obtain good scores in relation to the assessed methodological rigor, but due to their limited size, they provide insufficient evidence. Thus, it is crucial to further investigate how to obtain a fair assessment of the research quality, and how to distinguish between assessing the methodological rigor and the strength of evidence.

1.3 Research gaps and contributions

This thesis aims to contribute to the SE field investigating the different views of research quality in SE research. We intend to provide a foundation based on which we can facilitate a dialogue about quality of research and arrive at a shared view of what research quality means and how to assess the quality.

To achieve our main aim, we need to: (i) understand how measures of academic impact are related to quality of research; (ii) to define quality dimensions explicitly and be able to reason on their importance to research practice; and (iii) to provide the community with instruments to assess quality of research with regarding the quality dimensions.
1.3.1 Research Questions

In order to achieve our goals, our research comprises three main research questions:

RQ1. How are the research quality and academic impact of published research related?
Citation indexes have been used by the academic community as metrics of the impact of the research in the field. In an ideal scenario, this academic impact should also reflect the quality of the study regarding its methodological correctness, the relevance of provided evidence, and other standards for research quality. Assuming that academic impact and research quality are related, we aim to investigate how this relationship is accomplished.
To answer this research question, we compared scientometrics data with quality indicators from research reports that have been already assessed by the community. This is relevant to the main objective since it provides tangible evidence about existing bonds or gaps between the research practice and the conceptual dimensions of quality of research.

RQ2. Which quality standards are most important for software engineering research?
Although a common view of research quality has not been achieved, we identified candidate models proposing a comprehensive set of standards describing research quality. It is vital to assess whether such conceptual models describe the relevant aspects for the SE research field, and to understand the importance of each quality standard to decide which ones should inform software engineering research more than others.
To answer this research question, we gathered and compared perspectives from different researchers, assessing their level of alignment. This contributes to our main objective by eliciting the importance of the quality dimensions to decide which ones should inform SE research community more than others.

RQ3. How should we assess the quality of different types of empirical research?
Research appraisal is an essential task to ensure the quality of the produced research. It is often supported by assessment instruments, such as rubrics and checklist, which realize the quality standards into practices. Thus, by auditing, i.e., if the recommended practices are properly
conducted and reported, the review should ideally determine the quality of the study. There is a clear need to providing tangible evidence as a complement to the existing quality assessment instruments. These instruments should, in turn, be evaluated with regard to their applicability and comprehensiveness.

To answer this research question, we assessed available instruments and compared them to the quality standards considered important by the SE research community (see RQ2 above). We also derived a checklist to assess the quality of survey research as a complement to the existing assessment instruments (i.e., for case study and experiment). Finally, this contributes to our main objective by providing the research community with evaluated instruments to fairly assess empirical studies regarding their methodological rigor, relevance to practice and other dimensions of research quality.

1.3.2 Research Contributions

This thesis aims to contribute to the ESE field by providing a foundation to facilitate a dialogue about research quality and elicit the views of researchers about research quality standards and how to assess the quality. Also, we provided the following contributions, as illustrated in Figure 1.1:

C1. Make explicit the gap between research quality and academic impact. We investigated the relation between two quality dimensions (rigor and relevance) and scientometrics. The assessment scores were gathered from a well-used instrument to assess ESE studies. Results suggest that the assessed instrument does not relate to the academic community’s view of quality. Our findings motivate further investigation on how to acknowledge research with regarding the conceptual views of quality, and how to commit to such view by following appropriate guidelines.

C2. Synthesize the methodological support for SE research. We have identified and classified methodological guidelines and assessment instruments for ESE research. Despite the fact that a plethora of guidelines were found, assessment instruments for several methods (e.g., surveys) are lacking. Other methods (such as case study and experiment) are supported by assessment instruments with distinct views of quality. It is not explicit whether the assessment instruments are aligned to the views of importance of research quality to SE community.
Figure 1.1: Overview of the thesis contributions. The large rounded blocks describe the main components of the thesis, while the small boxes list the individual contributions and their rationale.
C3. Describe research quality and its components. We have identified a multidisciplinary model of quality dimensions [27], which aims to provide a shared understanding of research quality across different domains. Next, we gathered expert opinions from SE research practitioners regarding the relevance of that model’s quality dimensions to SE research. We incorporate the opinions of the participants into a revised model of research quality for ESE and detail a process to determine the importance of quality criteria using such a model. The result from the process aims for operationalization of quality dimensions as an assessment instrument.

C4. Provide solutions for the quality assessment of SE research. Based on the gaps identified in C2, we proposed and validated an instrument to assess the quality of survey research. The instrument is systematically derived from methodological support for SE research. We also validated the application of two checklist instruments to assess the quality of the experiment and case study research.

1.3.3 Research Methods
This thesis comprises a set of individual studies that collectively address the main objective. Therefore, a mixed-methods approach was used to answer the research questions, comprising a case study, survey, and focus group. A brief introduction to the different research methods used in this thesis is provided next; an overview of the different methods along with the contributions of individual studies is depicted in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Chapters</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic Mapping</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case Study</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focus Group</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimentation</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thematic Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>
Systematic mapping

An initial step towards larger research proposals comprises the investigation of literature that addresses the topic of interest. Systematic mappings make use of a structured process to identify and aggregate the existing literature and provide a comprehensive summary of the findings [23,30].

In chapter 3, we detail the process used to aggregate a catalog of methodological publications that provide support to ESE research. The results of this are further used as the basis for our studies investigating the checklist instruments (i.e. Chapters 5 and 6), which required the information provided by those identified publications. We also employed a systematic search strategy to collect relevant primary studies for our exploratory investigation in Chapter 2.

Case study

Case study research is used to investigate a phenomenon in its natural context, employing diverse methods of data collection and triangulation to provide a deeper understanding [2,34]. This methodology is more suitable for exploratory studies, providing a ground for further research.

We employed a case study approach to determine the alignment of the importance of a set of research quality dimensions. We made use of the Assessment Method for Software Quality (SAAM-SQ) [1] to collect the relative importance values, later comparing them to the opinions of SE researchers.

Focus group

Focus group is a research approach similar to interviews, but comprising a group of participants [24]. The focus group relies on the interactions among the participants to gather deeper qualitative evidence of the investigated phenomena. The researchers act mostly as moderators of the discussions and as observers of the group dynamics.

We used a structured focus group to collect the opinions of experts regarding a set of dimensions of research quality (Chapter 4). The data collected was later transcribed and analyzed through qualitative synthesis. We also used the focus group process in combination with the case study.

Experimentation

Experiments provide a more formal and rigorous method to investigate the relationship between factors of interest [38]. Experiments are used either as
1.3 Research gaps and contributions

explorative or evaluation objective, depending on the aim of the studies. Quantitative measures of the factors are collected during the experiment and further analyzed through statistical methods. We employed an experimental design to evaluate the performance of quality assessment instruments in a pedagogical context (Chapter 5).

Survey

Survey research is used to collect and summarize the opinions of participants regarding the topic of interest. It is vital that the sample of participants is aligned with the phenomenon of interest [32]. Later, the data collected can be synthesized either through quantitative or qualitative analyzed methods or even using a mixed approach.

We employed a mixed approach [22] to evaluate our proposed survey assessment instrument within the research practice context (Chapter 5). Our checklist was applied over a set of papers, and we surveyed the corresponding authors with regarding the completeness and fairness of our assessment.

Statistical analysis

Statistical methods are used to synthesize and interpret quantitative data, such as the one collected in an experiment or quantitative survey. It can be used to evaluate the strength of a relationship between factors (inferential statistics) or to better describe the effects of the phenomenon of interest (descriptive statistics).

In our studies, we employed both descriptive (i.e., clustering, statistical data distribution, statistical graphics) and inferential statistics (i.e., correlation analysis, conditional inference tree, inter-rater agreement) to analyze data. A more detailed description of such techniques is provided in Chapters 2, 4 and 5.

Thematic Analysis

Thematic analysis is a data analysis technique that is intended to provide an understanding of qualitative evidence, such as interview transcripts, documents, reports and other textual data sources [6]. Such an analysis method makes use of categorization through text coding to identify relevant themes. We employed a thematic analysis to construct our survey assessment instrument (Chapter 6).
1.4 Overview of chapters

This section provides an overview of each chapter in this thesis.

1.4.1 Chapter 2

While citations are utilized to measure the impact of research, the importance of achieving high-quality studies has been highlighted by the research community. One open question is whether study quality represented by rigorous methodology and practical relevance is related to academic impact, which would be desired. In this preliminary study we investigated how scoring rubric for rigor and relevance has been used to assess the quality of reported research in SE research; and explored the relationship between rigor, relevance and academic impact.

First, we selected a scoring rubric that covers both criteria highlighted as important (rigor and relevance) and has been used on a large set of studies. Through backward snowball sampling, we identified 682 primary studies using the scoring rubric. Hence, the basis for assessment was the same across all 682 studies. We utilized cluster analysis and conditional inference trees to explore the relationship between quality (here represented by rigor and relevance scores) and academic impact (here represented by normalized citations).

The results point out that the scoring rubric is not applied consistently. The rubric was found to be too abstract and needs to be further refined. Though, it can serve as an entry check before doing a more detailed quality assessment. The results showed that only rigor is related to the studies’ normalized citations. Also, other confounding factors likely influence the resulting citation impact. Our findings are further used as motivation to understand the relation between research quality and academic impact, and to discuss how to fairly acknowledge research studies for performing well concerning the emphasized quality.

1.4.2 Chapter 3

Some empirical studies have been carried out to date in SE, and the need for guidelines for conducting and evaluating such research has been stressed. In particular, guidelines and other methodological support intend to ensure that the research process is carried out according to quality standards. This chapter presents a mapping study aiming to identify and summarize the body of knowledge on methodological support on how to conduct and evaluate empirical research in SE.
1.4 Overview of chapters

Our systematic mapping process employed a manual search strategy and snow-balling search techniques to identify 341 papers reporting guidelines, assessment instruments and knowledge organization systems (KOS). To build up a catalog that research can use to identify relevant guidelines for their research problem, we extracted and categorized information provided by the identified papers. The methodological papers were classified according to research methods, research phases covered, and type of instrument provided. Later, we derived a brief explanatory review of the instruments provided for each of the research methods.

The catalog provides the SE research community with an aggregated body of knowledge on state of the art relating to guidelines, assessment instruments and KOS for carrying out empirical software engineering research. Our study identified potential opportunities for the development of SE research methodology. Explicitly, gaps for missing instruments to assess the quality of several research methods, e.g. survey. The catalog itself form a basis for identifying sources to derive such assessment instruments.

The chapter also discusses the catalog’s implications for research practice and the needs for further research. It is crucial for the next stages of our thesis work, in which this methodological support is used as a basis to address our research questions. We also present an exemplary usage scenario to guide those searching for a suitable set of methodological support literature, especially novices carrying out their first ESE studies.

1.4.3 Chapter 4

Research quality standards are intended to guide the execution and to assess the design and reporting of empirical studies. It comprises a series of concepts such as methodological rigor, practical relevance, and conformance to ethical standards. Depending on the perspective, different views of importance are given to the conceptual dimensions of research quality. Aiming to understand better what constitutes research quality from the perspective of the ESE research community, we assessed the level of alignment between researchers regarding a conceptual model of research quality.

In this chapter, we present a mixed methods research approach comprising an internal case study and a complementary focus group. We collected relative values for importance through hierarchical voting prioritization based on a conceptual model of research quality. Later, we gathered the participants’ opinions regarding research quality. Through focus group sessions, we also moderate
discussions with experts to address potential misalignment on their individual opinions.

The outcomes point out levels of alignment with regarding the importance of quality standards in the perspective of the participants. According to them, the conceptual model we presented fairly expresses the notions of research quality but has limitations with regards the structure and description of its components. Based on these results, we revised the conceptual model and provided an updated version adjusted to the context of ESE research. Later, we describe an exemplary scenario on how to use the revised model to characterize a quality assessment instrument. We also discussed how to employ our research approach to assess quality alignment in research.

1.4.4 Chapter 5

On the context of higher education, it is important teaching students to appraise scientific literature critically. This is the case for a postgraduate research methodology course we investigate in this chapter. The course uses checklists for assessing the scientific rigor of empirical studies, thus supporting students in reviewing the scientific literature.

We employed an experimental design where 76 students (in pairs) used two checklists to evaluate two ESE papers (reporting case studies and experiments) each. We compared the consistency of students' assessments within themselves, and the accuracy of their ratings against the ones from more senior researchers, with varied results. We also collected the students' perception of using checklists. Checklist items related to data analysis are rated as difficult to assess.

A factor seemed to be that the clearer the reporting, the easier it is for students to judge the quality of studies. Our results reinforce the needs for communicable scientific literature, as it is important that authors write to enable synthesis and quality assessment. The results also support the continued use of checklist instruments as a means for introducing scientific reviews.

1.4.5 Chapter 6

Checklists have been used to assess the methodological quality of different types of empirical studies, especially experiments and case studies. However, assessment instruments for other research methods (e.g., surveys) are lacking. Over the past decade, a steady increase in survey-based studies has been noted, and there are several guidelines providing support for those willing to carry out surveys.
1.5 Synthesis/Discussion

Thus, this chapter aims to address the identified gap, proposing and evaluating a checklist to support the design and assessment of survey-based research in SE. To construct the checklist, we systematically aggregated knowledge from 14 methodological papers supporting survey-based research in software engineering. We identified the key stages of the survey process and its recommended practices through thematic analysis and vote counting. Further, we evaluated the checklist in the research practice context.

The evaluation provided insights regarding the limitations of the proposed checklist in relation to its understanding and objectivity. Specifically, 19 of the 38 checklist items were improved according to the feedback received from experienced researchers. Finally, the chapter promotes a discussion on how to use the checklist for auditing surveys reports as well as a support tool to guide ongoing research about the survey design process.

1.5 Synthesis/Discussion

This section discusses how each of the chapters addresses our research questions and provides a contribution to the field.

1.5.1 Research Methods and Tools

The emphasis of the thesis lies in the methodological and practical aspects of empirical software engineering research. An overview of the individual components that are part of the thesis and their positioning concerning the main contributions can be found in Section Figure 1.1.

Initially, we investigated whether the assumption that quality of research was not regarded as a measure of impact was right (Chapter 2). We grounded this assumption on the discussions of experts (e.g., [11, 29]) and further collect evidence from the field that could support it. The core of our investigation is a potential relationship between scientometrics and the quality assessment of SE research publications.

The results point out that practical relevance is not particularly regarded with academic impact. Although practical relevance aims an impact on the industrial context, the academic community is also expected to acknowledge such contribution, and its potential to guide future original research. The evidence also shows a weak relationship between the academic impact of studies and their methodological rigor.
Rigorous research relies on the correct application of the research methodology to explore or explain a phenomenon of interest. The validity of the study and trustfulness of produced evidence depend on these methodological standards. To successfully investigate such aspects, one should first become familiar with the different research methods, their strengths, and weaknesses.

We have compiled a comprehensive set of guidelines and other supporting literature for empirical research in SE (Chapter 3). To produce such a catalog, we investigate the state-of-art by systematically mapping 341 methodological studies. We also derived a set of heuristics for researchers willing to use the catalog to find a set of relevant papers to their proposed research.

Ultimately, we expect to impact the quality of future ESE research, as researchers using appropriate guidelines produce more rigorous empirical studies. We believe that research practice results are strongly dependent on research’s methodological quality. Thus, by providing researchers with evaluated methods and practices, we are fostering more credible, relevant evidence. This philosophy guided the intent of this thesis’ work and has also been the foundation for the research practice presented herein.

1.5.2 Views of Research Quality

We address the central aspect of this thesis through two distinct viewpoints (illustrated in Figure 1.2). The first one is that we employed a top-down approach to gathering perspectives from researchers regarding concepts of research quality. The second one is that we investigated the appropriateness of instruments to assess the quality of reported studies are, i.e., how well they perform according to the needs of research practice.

The top-down approach considers a set of research quality dimensions proposed in Mårtensson’s et al. [27] conceptual model. We ought to examine whether this multidisciplinary model fairly represents research quality in the views of research practitioners, and by doing so, to understand better what is valued as quality in SE research. The study described in Chapter 4 used a mix-methods approach to measure how well the views of different researchers regarding the relative importance of the quality dimension align. Grounded in this importance metrics, we fostered discussions that lead to deeper reasoning of research quality concepts.

We do not assume to have achieved the highly desired shared view on the quality of research. However, we identified interesting findings, such as (i) researchers acknowledge the completeness of the dimensions in the conceptual model but does not agree with the structure and description of its components;
and (ii) there is a need to detail how to operationalize the model, in order to assess whether it is appropriate.

The main contribution of this chapter is to provide reasons for the alignment and misalignment of SE researchers regarding a common view of research quality. The suggestions of participants were incorporated into an evolved model that, we suppose, is a step closer to shared understanding. It is essential to validate this view and to disseminate among the research community. Our findings suggest that further work is needed for a continuous evolving of such a conceptual model. Aiming that, we also detailed the approach we employed to assess the alignment of importance among participants and discuss how the research community can use it.

Figure 1.2: The double approach to investigate the main focus of this Thesis.
The bottom-up approach focuses on the operationalization of the quality assessment of ESE studies using different research methods. In this context, we investigated how to employ checklists to appraise scientific publications (Chapter 5). In particular, we focus on checklists prescribed for case study research [16] and experiment [38] in SE; and later complement with our proposal for the survey method (Chapter 6).

An experimental study conducted with students from the Research Methodology course provide insights regarding the consistency of ratings and applicability of the instruments. Different from Kitchenham et al. [20], our participants have limited experience with the scientific research appraisal, so we also intended to investigate the pedagogical potential of using the checklist in education.

Our results regarding the accuracy are poor even when employing joint reviewers. Despite that, students performed better with regarding high-quality publications. The student’s reckoned that the checklists are not difficult to answer, but the assessment depends on the structure and contents of the publication and the reviewers’ knowledge. Implications for a pedagogical approach using checklists to foster critical thinking and analytical skills are provided. Moreover, the recommendations are also valid for more experienced reviewers, and to researchers proposing or evaluating such assessment instruments.

In the study described in Chapter 3, we identified instruments to assess the quality of SE research. Checklists for case study research and experiments are not unheard of, but this is not the case for other research methodologies, e.g. survey. We further address this gap in Chapter 6, by designing a checklist for survey-based research and carrying out an evaluation both in the context of research practice.

The checklist is composed of recommended practices aggregated from a set of survey-focused methodological SE papers. The evaluation employed a survey-based design to gather data on perceptions from using our proposed checklist to assess survey-based research. Overall, the participants (i.e., researchers that have conducted survey research) considered our assessment tool fair and comprehensive. Their feedback was further analyzed in order to improve the assessment instrument.

Possible implications of the checklist to guide the conduction and to audit survey-based research are discussed. Notably, we detail decision-making points in the process and the potential implications of the decisions to further stages of the survey process. We also suggest researchers employing the checklist to guide their research to reflect on such decisions and provide a summary in the publication. Overall, the results from our studies suggest that a fairer quality assessment depends on providing proper information regarding the research.
1.6 Conclusion and Future work

In this thesis, we investigated different views of the research quality of empirical studies in software engineering. In particular, we involved researchers in discussion regarding quality standards for ESE research, why and how to adopt them. This discussion aims to contribute to the development of a more evidence-based approach to software engineering, by recommending techniques and methods to assess the quality of research.

Based on the findings of this thesis, the following conclusions are drawn:

**RQ1. How are the research quality and academic impact of published research related?**

In order to evolve the knowledge on the field, SE researchers are encouraged to design and carry out studies according to standards of excellence, especially methodological rigor and practical relevance [17, 26]. However, the researchers' performance is evaluated through citation indexes, which only their impact on academia [29]. The citation count of a paper is influenced by a series of factors (e.g., availability of the report, the structure of the document), and it is not possible to isolate the effects of research quality alone.

SE research community acknowledge the efforts to ensure strictness of research, i.e., rigorous studies are more likely to be published and cited. This seems not the case for practical relevance. We exposed a misalignment between the potential relevance and academic impact of ESE research. The potential relevance is rated using a set of rubrics [17] that assess the realism of the studies but does not reflect the actual impact of research to practice.

In addition to that, the assessment rubrics we use to investigate this relation are not employed consistently. In some cases, researchers modified the rubrics to adapt it to their needs, whether to answer research aims or to select studies according to their perceived quality. If the rubrics are not complete, there is a need to identify a set of criteria that better represent the quality standards desired by the ESE community, grounded on the rationale that guide their use.

**RQ2. Which quality standards are most important for software engineering research?**

Excellence on empirical research depends on the methodological rigor and ideally on the practical relevance [2, 8]. Despite these primary values, holistic models of research quality also account the importance of, e.g,
good science communication and legitimacy [27, 28]. Also, each of such standards comprises a set of subdimensions not likely to be equally valued by researchers.

There is hardly an aligned view of importance with regarding quality dimensions for research. In our study, experts agreed more about some dimensions than others, particularly the methodological strictness, trusted evidence, and practical relevance are favored. This scenario is not likely a fair representation of most important quality standards to the field, as individual perspectives could change due to the objective and characteristics of each intended study. The misalignment on the views of importance we observed is likely to happen in different contexts.

The methodological approach we used showed itself suitable for fostering the discussion and elicit the alignment among researchers and experts. Possible ways to employ such an approach to evolve the conceptual research model have been proposed.

RQ3. How should we assess the quality of different types of empirical research?

Reasons to assess the quality of research are manifold [28]. Researchers themselves should ensure their research practice meet the standards for excellence. The peer-review process employs quality assurance as a mean to audit research worth to be published. Either way, it is vital to operationalize models of research quality, producing instruments able to guide and audit research.

Assessment instruments can be systematically derived from well-established research methodologies. This approach has the potential to link the two perspectives we discuss above, i.e. research output is assessed through the same rigor employed during the conduction of the study. Based on these principles, we have proposed and validated a checklist to guide and assess survey-based research. The instrument is intended to fill a gap regarding the assessment of empirical studies in SE.

Similar checklists exist for reviewing the case study and experimental research. Such instruments were validated in different contexts and also employed in the educational context, to instruct novice researchers to appraise the quality of published research critically. Such approaches to promote critical thinking are essential to the development of future ESE research practice, as the ones pursuing careers as researchers become more critical about the quality of studies they will produce and review.

Having reached our objectives, we hope to encourage SE research commu-
nity to strive for excellence in research practice, not only in the rigorous application of the research methods and tangible evidence on research results but also in considering a more holistic view of research quality. Such contribution has a potential for the maturity of SE research field, as a substantial body of high-quality empirical knowledge is produced and evaluated by academics and applied by practitioners in industry.

1.6.1 Future Work

Finally, we recognize that the work is not entirely done. Our highly desired objective of providing an SE research community with a shared view of what constitutes research quality is still preliminary. It is vital to mature and evolve such view, disseminating the knowledge and aggregating the perceptions and reasoning from the overall community.

Our top-down and bottom-up approaches address the research quality through complementary views, i.e. conceptual and instrumental. However, the consolidation of the two views into a complete paradigm remains to be completed. Further works towards this objective are:

Survey to investigate experts’ perceptions regarding the model of research quality. Our focus group approach made explicit a series of misalignments regarding the dimensions of research quality. Still, the participants of the workshops at the ISERN annual meeting are a small parcel of the academic community interested in methodological support for research. It would be beneficial to collect and aggregate the perceptions of a large number of experts.

In Chapter 3, we aggregated a catalog of methodological support for 16 research methods in ESE. The authors of such methodological papers are the target population for a survey study investigating the notions of research quality. We ought to present to them the set of dimensions of Mårtensson et al. [27] conceptual model and survey whether the dimensions are relevant and feasible to operationalization in the SE context.

Finally, we plan to triangulate the results from our planned survey with the results from the focus group workshops. By doing so, we can evaluate the strength of alignments and misalignments, and thus better assess the recommendations to be done in order to evolve the conceptual framework.

Comparison and aggregation of checklists to assess case studies, experiments, and survey. We are currently investigating how the currently available solutions for quality assessment of different research methods (e.g., case study, experiment, and survey) compare. Our goal is to provide a shared instrument aligned to perceptions of the ESE community obtained in Chapter 4. To achieve this
objective, we have collected and categorized the checklist items using thematic analysis, using the revised model of research quality as a preliminary codebook.

Further, we plan to summarize the themes in order to identify checklist items that to be considered regardless the methodology; so, to compose a shared instrument; and checklist items that are particular to a research method, so denoting extensions to the shared instrument. It would also be beneficial to summarize themes that represent dimensions not expressed in the conceptual model of research quality, thus potential candidates to continue evolving it.

There is also a potential implication to the pedagogical context described in Chapter 5, as a tailored checklist has the potential to instruct novices on how to assess and compare the quality of different studies regardless of the research approach. A common assessment could also benefit the aggregation of knowledge, as such as the produced by literature reviews and meta-analysis.

1.7 References


1.7 References


28 Introduction


1.7 References


Chapter 2

Towards understanding the relation between citations and research quality in software engineering studies

This chapter is based on the following paper:


2.1 Introduction

In an ideal scientific world, we expect research reports (e.g., academic papers) to be of high quality, to present a significant contribution to the body of knowledge, and also to have their impact reflected in the number of citations [42]. Such scientific reports should provide the research community with new insights, describe rich cases and experiences, propose new methods or evaluate those already established. Despite differences between the different scientific disciplines, this reflects the expectation on most research fields.
Towards understanding the relation between citations and research quality in software engineering studies

For this aim, quality in the reporting of research has been investigated in a multidisciplinary context (see, e.g. [47, 77]), where several dimensions (such as credibility, contribution, communication, and conform) were reported as quality standards for scientific research. These standards reflect recommendations and best practices provided in guidelines and supporting literature; they are also the basis for instruments assessing reported studies, thus evaluating the potential impact to practice.

However, aspects other than the quality dimensions may have a stronger relation with impact factors, positively or negatively influencing them. For example, paper characteristics, such as number of authors, publication type and venue; and research aspects, such as discipline and industrial applicability, are candidates to present such an effect upon scientific impact [41, 44].

Therefore, we believe that it is important to understand more precisely how quality dimensions (herein rigor and relevance) have been evaluated and how they relate to one another, and to scientific impact. Such knowledge can contribute as an effective screening and selection criterion to guide researchers and practitioners in relation to which studies they should read and rely upon, and also to encourage them to be used as a yardstick to assess existing research fairly and accurately.

Given the need to understand how the quality standards we are using today relate to scientific impact, we conducted an exploratory study aimed to explore the reporting of research with respect to rigor and relevance as well as its relation to scientific impact in software engineering. In particular, the research presented in this Chapter makes the following contributions:

C1) Provides an overview on how Ivarsson and Gorschek’s [70] scoring rubric has been used to assess the quality of primary studies in systematic literature reviews (SLRs) and systematic mapping (SM) studies. Researchers may find this assessment useful to further develop and apply the rubric for assessment of primary studies. Furthermore, potential limitations of the rubric will become apparent.

C2) Identifies existing relations between scientific impact factors and the assessed research quality in software engineering (SE) studies. Researchers may utilize the findings as a reflective tool for which studies to include in their related works. Furthermore, the findings may open a discussion and reflection on the current and desired relation between rigor, relevance and scientific impact.
2.2 Background and Related Work

The remaining of the Chapter is organized as follows: We summarize related concepts in Section 2.2, and describe the methods we used to obtain the data and to conduct our empirical analysis in Section 2.3. Section 3.4 presents the particular findings for each investigated objective. Later, we discuss the results in Section 2.5 and point out some of the limitations of our work in Section 2.6. Finally, in Section 3.6, we conclude our study and provide suggestions for future work.

2.2 Background and Related Work

To better understand the context in which this exploratory study is situated, we describe the related literature of three relevant themes: 1) background of empirical research in software engineering and its related challenges; 2) quality assessment of research, represented by rigor and relevance as dimensions; and 3) bibliometric analysis, as well as metrics and methods to evaluate scientific impact.

2.2.1 Empirical Research in Software Engineering

Software engineering (SE) is an emerging and in maturation sub-discipline of computing. It makes use of scientific and technical knowledge to the development and maintenance of software. SE research is mostly aimed at industrial practice, in particular, to supply practitioners with means to better decision making on whether or not to adopt software technologies and development methods [72,100].

Similarly to other applied research fields, SE builds its body of knowledge upon meaningful evidence. This evidenced-based approach employs empirical research methods (such as experiments, cases studies, and surveys) to evaluate methods, techniques and tools [72, 100]. Further, guidelines have also been proposed to support the research process (see, e.g. [86, 100]) and assess the resulting evidence (e.g. [66, 70]).

However, like several other fields of science and engineering, empirical software engineering (ESE) has not yet developed a well-established method to evaluate the contribution of research practice [77, 88]. While academic impact is often measured through citation counting [40], the level of industrial impact lacks feasible standards. Notwithstanding, one expects high quality studies to produce stronger evidence, thus potentially more relevant to industry practice.
Towards understanding the relation between citations and research quality in software engineering studies

2.2.2 Assessing Research Quality

Quality assessment is an important activity in research practice, as it ensures that the results of the assessed studies are meaningful, i.e. provide strong evidence [53]. The quality in the reporting of research relates to the completeness of information needed to judge the study according to standards. The quality assessment process requires the use of an instrument [71], and its results should support the reported findings and/or identify threats to the study’s validity.

Along these lines, rigor and relevance are two perspectives of quality that address how the reported research contributes to the body of knowledge, and its potential to transfer knowledge from research to practice [47]. The two dimensions have been further investigated across several different domains, such as business, psychology, and social sciences (e.g. [45, 68, 92]).

On the one hand, rigor is usually emphasized more in academic than in industry environments, as it refers to the precision and correctness with which a study is reported regarding the research method used [47]. To evaluate rigorousness, quality concepts such as internal validity, reliability and, contextuality are assessed according to the study design [76, 77].

Relevance, on the other hand, represents the study’s potential usefulness in the target context. Relevant research is focused on address problems and on providing real value to practitioners [47]. Despite this, evidence from several studies advocate a lack of industry interest on scientific research, e.g. [56, 102]. One of the possible causes is the distance between academic research objectives and industry demands [85]. This gap can be abridged by conducting research in an environment that closely resembles the context in which it is intended to benefit. Quality concepts such as interesting and current idea, applicable results, and accessible presentation [47, 76] are related to relevance.

Although rigor and relevance are not the only quality dimensions related to research practice, their importance has been further stated in [47, 76, 78]. Ivarsson & Gorschek’s scoring rubrics [70] particularly assess the extent to which aspects related to rigor (as summarized in Table 2.1), and the potential for impacting the industry (described in detail in Table 2.2) are reported.

2.2.3 Scientometrics

Scientometrics refers to the study of measurement aspects such as performance, impact, international collaboration, etc. of scientific and technological activities based, in particular, on citation analysis [95]. Citations are used to rank scientific journals, papers, research organizations and individuals as follows [40]:
Table 2.1: Scoring rubric for evaluating rigor [70]

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Strong description (1)</th>
<th>Medium desc. (0.5)</th>
<th>Weak desc. (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context (C)</td>
<td>The context is described to the degree where a reader can understand and compare it to another context. This involves description of development mode, e.g., contract driven, market driven etc., development speed, e.g., short time to market, company maturity, e.g., start-up, market leader etc.</td>
<td>The context in which the study is performed is mentioned or presented in brief but not described to the degree to which a reader can understand and compare it to another context.</td>
<td>There appears to be no description of the context in which the evaluation is performed.</td>
</tr>
<tr>
<td>Study Design (SD)</td>
<td>The study design is described to the degree where a reader can understand, e.g., the variables measured, the control used, the treatments, the selection/sampling used etc.</td>
<td>The study design is briefly described, e.g., “ten students did step 1, step 2 and step 3”</td>
<td>There appears to be no description of the design of the presented evaluation.</td>
</tr>
<tr>
<td>Validity Threats (V)</td>
<td>The validity of the evaluation is discussed in detail where threats are described and measures to limit them are detailed. This also includes presenting different types of threats to validity, e.g., conclusion, internal, external and construct.</td>
<td>The validity of the study is mentioned but not described in detail.</td>
<td>There appears to be no description of any threats to validity of the evaluation.</td>
</tr>
</tbody>
</table>
Table 2.2: Scoring rubric for evaluating relevance [70]

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Contribute to relevance (1)</th>
<th>Do not contribute to relevance (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users/Subjects</td>
<td>The subjects used in the evaluation are representative of the intended users of the technology, i.e., industry professionals.</td>
<td>The subjects used in the evaluation are not representative of the envisioned users of the technology (practitioners). Subjects included on this level is: i) Students, ii) Researchers, and iii) Subject not mentioned.</td>
</tr>
<tr>
<td>(U)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context (C)</td>
<td>The evaluation is performed in a setting representative of the intended usage setting, i.e., industrial setting.</td>
<td>The evaluation is performed in a laboratory situation or other setting not representative of a real usage situation.</td>
</tr>
<tr>
<td>Scale (S)</td>
<td>The scale of the applications used in the evaluation is of realistic size, i.e., the applications are of industrial scale.</td>
<td>The evaluation is performed using applications of unrealistic size. Applications considered on this level is: i) Down-scaled industrial, and ii) Toy example.</td>
</tr>
<tr>
<td>Research method (RM)</td>
<td>The research method mentioned to be used in the evaluation is one that facilitates investigating real situations and that is relevant for practitioners. Research methods that are classified as contributing to relevance are: i) Action research, ii) Lessons learned, iii) Case study, iv) Field study, v) Interview, and vi) Descriptive/exploratory survey.</td>
<td>The research method mentioned to be used in the evaluation does not lend itself to investigate real situations. Research methods classified as not contributing to relevance are: i) Conceptual analysis, ii) Conceptual analysis/mathematical, iii) Laboratory experiment (human subject), iv) Laboratory experiment (software), v) Other, and vi) N/A.</td>
</tr>
</tbody>
</table>
1. for journals, the impact factor (IF) is given by the average number of citations for a collection of articles published in preceding years (e.g., 5-year IF is the average number of citations in the past five years);

2. for individual papers, the number of citations is used as a measure of scientific impact; however, quite often the impact factor of the journals in which those papers are published ends up being the one used as an individual paper’s surrogate measure of scientific impact;

3. for individual researchers, the most commonly used research impact measure is the h-index, i.e. the largest n for which an individual has published n articles, each with at least n citations.

Citation count is important as an indication of the influence of scientific reports. Papers cited extensively often provide insights and experiences, describe new research directions, or summarize the state-of-the-art or practice in a specific field. Wohlin conducted a series of studies to analyze the most cited articles in Software Engineering journals through the years, comparing them to highlight the similarities and differences (see, e.g. [98]).

Similarly, other research studies investigated aspects related to scientometrics of SE publications. Two particular studies report a census of SE publications: the first characterizes the papers published in IEEE Transactions on Software Engineering journal in the period between 1980 and 2010 [64]; and the second highlights time-related trends of SE papers listed in the DBLP database\footnote{http://dblp.uni-trier.de} with publication dates between 1971 and 2012 [55]. In a different scope, the particular case of Turkish SE research is investigated in [60].

One could expect that, due to the proximity of researchers to the scientific community, measures of scientific impact (such as citation count) would be more strongly related to academic rigor than to industrial relevance. If a relationship between relevance to practice and scientific impact is not established, this could motivate the research community to investigate metrics for assessing industrial impact, i.e. how broadly academic research addresses industry demands [63,79].

Although a well-established metric, citation count faces several criticisms on evaluating scientific impact. In an article called “Stop the Numbers Game”, Parnas [80] suggests that the citation count slow down scientific progress, as it encourages researchers to publish several superficial papers rather than a few correct and relevant ones. This issue is further investigated in [62] by means of a quantitative bibliometrics assessment. The results show that particular paper
Towards understanding the relation between citations and research quality in software engineering studies

characteristics (such as publication venue and language) are likely to have higher impact than citation count.

Alternatives to the citation metrics have been proposed to more fairly evaluate the impact of research. Examples of those include a taxonomy to assess the citation behavior [48], and the use of citation distribution rather than a single point measure [99].

Furthermore, critics argue that the quality assessment should be emphasized over citation counting [42,80], but so far we only have a limited understanding of how the quality in the reporting of research is related to scientific impact. Investigations on the topic are usually focused on citation analysis of published studies [65,84,99,101].

2.3 Research Method

To investigate how reporting quality relates to scientific impact, we have first identified rigor and relevance as two relevant criteria for which data can be collected. Based on the gathered data, we employed statistical and visualization mechanisms to understand their relationship. The data used herein was obtained via a systematic search and collection of SLRs and SMs, the quality assessment scores for the primary studies they included, and additional paper characteristics.

2.3.1 Research Questions

This study explores two objectives related to reporting quality assessment of ESE research and its relation to the scientific impact. The objectives lead us to distinct contributions, as presented in Section 2.1. Each contribution is guided by its own research questions.

First contribution (C1) provides an overview of the scoring rubrics’ use. Therefore, we aim to investigate the insights of researchers using the rubric to assess the quality of reported empirical studies. Based on this, we formulated two research questions:

RQ1. How the scoring rubrics were applied? This question is focused on the methodological aspects of applying the instrument.

RQ2. What purpose the rubrics were used as/for? This aims to assess the reasons for scoring the included papers.
The second contribution (C2) explores the existing relations between impact factors and a study’s reporting quality. A statistical model was built to test the following hypothesis:

$H_0$: There is no significant relation between rigor and relevance criteria and the scientific impact of studies (i.e., normalized citations per year) in ESE research.

$H_A$: There is a significant relation between rigor and relevance criteria and the scientific impact of studies (i.e., normalized citations per year) in ESE research.

### 2.3.2 Construct Measures

The variables we aim to investigate are divided into two categories: the ones assessing the quality in the reporting of research, and the ones addressing its visibility [96]. In our work, research quality comes from the assessed rigor and relevance criteria, while visibility is represented by the scientific impact. For each of these, we identify candidate metrics or assessment instruments, listed in Table 2.3.

Rigor and relevance are often evaluated by a common assessment instrument. This is usually a checklist, in which the individual questions address a particular aspect or degree. The questions can be subjective and the aspects evaluated could overlap. Moreover, other quality dimensions (e.g., originality and credibility) are also addressed.

In this study, the main concerns are related to a comparable measure for both rigor and relevance. To achieve such an aim, a quantitative measurement scale for each dimension is desirable. Considering this constraint, Ivarsson & Gorschek’s rubrics [70] were selected. The scoring rubrics assess both reporting rigor and relevance of research in SE, as summarized in Tables 2.1 and 2.2. Ivarsson & Gorschek [70] also conducted a validation study by applying the proposed method to an SLR on requirements engineering.

Scientific impact is the product of citation analysis, and several measures have emerged to compute it. Despite criticism, counting the number of citations is the most straightforward method to measure such dimension [40]. It is important to consider the progress in the citation counts over the years since paper’s publication. During this life cycle, several factors can influence its scientific impact, such as the motivations for referencing the study [48,96] and the characteristics of the scientific communication process [96]. Therefore, other paper characteristics (e.g., type and venue of publication, research method used)
Towards understanding the relation between citations and research quality in software engineering studies

Table 2.3: Candidate measures for the investigated criteria. The highlighted rows point out the selected measures for the investigated variables (i.e. rigor and relevance, and impact).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Options</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rigor and Relevance</td>
<td>CASP Qualitative Checklist [39]</td>
<td>Address the rigor, credibility, and relevance issues through ten questions. The questions are not mapped to the quality dimensions. It was developed for Evidence-Based Medicine and is broadly applied.</td>
</tr>
<tr>
<td></td>
<td>Dybå &amp; Dingsøyri [52]</td>
<td>Address context, rigor, credibility, and relevance criteria. There is only one relevance question addressing the value provided for research or practice.</td>
</tr>
<tr>
<td></td>
<td>Ivarsson &amp; Gorschek [70] (selected)</td>
<td>Address rigor and relevance through 3 and 4 questions, respectively. Results are computed in an ordinal scale.</td>
</tr>
<tr>
<td>Impact</td>
<td>Absolute number of citations [40]</td>
<td>Not appropriate to compare papers with distinct ages (i.e., published in different years).</td>
</tr>
<tr>
<td></td>
<td>Average number of citations [40, 61] (selected)</td>
<td>Citations are not equally distributed over the years.</td>
</tr>
<tr>
<td></td>
<td>Impact factor [40]</td>
<td>Journal-level metric. Provides no information on specific paper.</td>
</tr>
</tbody>
</table>

can act as confounding factors for this metric. However, this exploratory study does not aim to explain the causes for the citations evolution.

Mature studies are more likely to achieve higher citation count, thus a normalized metric for citations is desirable to conduct comparative analysis. The average number of citations per year is a ratio measure commonly used to compare citation counts for papers between fields [61, 96]. It is calculated using the arithmetic mean, i.e. dividing the absolute number of citations that a paper obtained by the number of years since its publication [61].

2.3.3 Study Identification and Data Collection

First, we identified Ivarsson & Gorschek [70] as the starting paper for a reference-based search returning 55 papers. Among those, we identified 16 secondary studies (SLRs and SMs) using the scoring rubrics to assess included primary studies. The list of candidate papers is presented in Table 2.4. Further, we
assessed each of those candidate papers to select 12 of them providing rigor and relevance assessed scores (overall and individual scores for each scoring dimension). The process for identification and selection of the studies is illustrated in Figure 2.1.

Table 2.4: Candidate papers for the exploratory study.

<table>
<thead>
<tr>
<th>ID</th>
<th>Paper Type</th>
<th>Assessment Scores</th>
<th>Primary Studies</th>
<th>Data Origin</th>
<th>Missing scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 [43]</td>
<td>journal (JSS)</td>
<td>Detailed</td>
<td>87</td>
<td>Collected from the paper</td>
<td>Relevance: Context and Research Method</td>
</tr>
<tr>
<td>S2 [46]</td>
<td>journal (IST)</td>
<td>Detailed</td>
<td>43</td>
<td>Collected from the paper</td>
<td></td>
</tr>
<tr>
<td>S3 [50]</td>
<td>journal (JSS)</td>
<td>Detailed</td>
<td>58</td>
<td>Asked by mail</td>
<td>Values reported as N/A instead of 0 for Relevance sub-items: Users/Subject and Scale</td>
</tr>
<tr>
<td>S4 [51]</td>
<td>journal (IST)</td>
<td>no</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5 [54]</td>
<td>journal (IST)</td>
<td>no</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6 [57]</td>
<td>journal (TSE)</td>
<td>Detailed</td>
<td>196</td>
<td>Asked by mail</td>
<td>All (different methods mapped to rigor and relevance scores)</td>
</tr>
<tr>
<td>S7 [69]</td>
<td>master thesis</td>
<td>no</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8 [74]</td>
<td>journal (IST)</td>
<td>Detailed</td>
<td>46</td>
<td>Collected from the paper</td>
<td>All (different methods mapped to rigor and relevance scores)</td>
</tr>
<tr>
<td>S9 [78]</td>
<td>journal (IST)</td>
<td>Detailed</td>
<td>41</td>
<td>Collected from the paper</td>
<td></td>
</tr>
<tr>
<td>S10 [81]</td>
<td>journal (IST)</td>
<td>Detailed</td>
<td>43</td>
<td>Collected from the paper</td>
<td></td>
</tr>
<tr>
<td>S11 [82]</td>
<td>journal (JSS)</td>
<td>Detailed</td>
<td>38</td>
<td>Asked by mail</td>
<td></td>
</tr>
<tr>
<td>S12 [87]</td>
<td>master thesis</td>
<td>no</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S13 [93]</td>
<td>master thesis</td>
<td>Overall scores only</td>
<td>41</td>
<td>Collected from the thesis</td>
<td></td>
</tr>
<tr>
<td>S14 [94]</td>
<td>master thesis</td>
<td>Detailed</td>
<td>89</td>
<td>Collected from the thesis</td>
<td></td>
</tr>
<tr>
<td>S15 [59]</td>
<td>journal (CLEIej)</td>
<td>Detailed</td>
<td>18</td>
<td>Collected from the paper</td>
<td></td>
</tr>
<tr>
<td>S16 [58]</td>
<td>Ph.D. thesis</td>
<td>Detailed</td>
<td>18</td>
<td>Collected from the thesis</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>12</td>
<td>718</td>
<td></td>
</tr>
</tbody>
</table>

Further, we gathered the references for each of the primary studies assessed by the papers in Table 2.4 (totaling 718 primary studies) were stored in a dataset and is characterized by the following variables: the assessed scores for rigor and relevance (and their individual aspects), the study impact (total number of
Towards understanding the relation between citations and research quality in software engineering studies

Figure 2.1: Process for identification and selection of candidate studies.

- citations, year of publication, and normalized citations per year) and additional paper characteristics (type of the paper, publication venue, and length in pages). These characteristics were collected as possible confounding factors so their influence could be tested/evaluated and also discussed inside the SE domain (see, e.g. [96]). To enable a critical examination of our dataset and to facilitate further studies, we have made our dataset available at https://goo.gl/3y7R4l.

As this exploratory study was carried out between mid and late 2015, we computed the normalized citation counts obtained up to 2014. This means that the most recent studies (i.e., from 2013) had at least one year of citations covered. The gathered dataset (see Figure 2.2) covers both novel and mature primary studies, with citation counts varying from 0 to 108.

2.3.4 Analysis

We examined the dataset and checked whether there was a significant amount of detailed scores and missing values, as summarized in Table 2.4. From the 16 identified studies, 7 does not provided enough information in the paper. We then contacted the authors by e-mail kindly asking them to share the list of included studies along with the scores for rigor and relevance. Three authors answered our request, thus increasing our dataset to 12 included studies. The remaining 4 studies were excluded due to the lack of detailed scores to perform the data analysis.

With regard to the first contribution (C1), some of the selected studies adjusted the scoring rubrics by addressing individual aspects in a different manner:

S1 Does not assess two sub-aspects related to relevance, i.e. context (C) and research method (RM). These two aspects are part of the search strategy,
Figure 2.2: Distribution of the dataset according to the number of primary studies published in each year. The shading segments of the columns represent the normalized citation counts (cit/year), i.e. a stronger shade means a higher number of citations per year, whereas the lighter ones show the less cited papers. A legend on the right side shows a sample of the shades within the range of minimum and maximum citation counts, zero (0) and 108, respectively.
thus all included studies fulfill both criteria. Therefore, we replaced the missing scores with a contribution (1) score;

S3 Instead of the null (0) value, the study reports some instances as non-applicable (N/A). This is mostly opposed to the rubrics’ proposed use, whereas null represents the no fulfillment of a particular rigor or relevance criterion. In this case, we replaced the N/A values for a null (0) score; and

S6 and S8 Both studies map another instrument [52] to assess rigor, and used evidence levels as relevance scores. To collect the rigor dimension scores, we aligned each checklist item to Ivarsson and Gorschek’s [70] rubrics. For relevance, we normalized the evidence levels (ordinal data ranging from 1 to 6) to overall relevance range (from 0 to 4). This normalization resulted in relevance scores dissimilar to those expected for the original rubrics (e.g., 0.8 instead of 1) but which are still suitable as an ordinal independent variable.

Further information regarding the application and refinement of the rubrics is given in Section C1, in which we discuss the first contribution of this study: an overview on how the scoring rubrics have been used to assess the reporting quality of primary studies in SLRs.

Relationship Inference

The second contribution (C2) of this study is achieved by exploring the relation between rigor and relevance scores of primary studies (i.e., independent variable) and the computed citations per year for each of these primary studies (i.e., dependent variable). Based on the contribution proposed in Section 2.1, we test the hypothesis presented in Section 2.3.1. A series of statistical approaches are conducted to investigate the relation between reporting quality (i.e. rigor and relevance) and scientific impact, as follows:

Correlation analysis: A preliminary investigation of the relationship is made by applying correlation analysis to the dataset. The approach measures the extent of statistical covariance, i.e. which two observed variables tend to change together [75]. Spearman’s correlation coefficient (also known as Spearman’s $\rho$ or $r_s$) rates the degree of linear dependence between two variables, describing both the strength and the direction of the statistical relationship. We opted for this particular correlation coefficient because its rank approach is less subjective to particular distribution assumptions.
Clustering variables: A statistical approach to detect subsets of strongly correlated variables, i.e. which provide the same information, or belonging to a common group. The approach is especially useful to identify underlying structures and redundancies between variables for dimension reduction treatments [49]. We used a hierarchical agglomerative method for data partition, iteratively aggregating the less dissimilar clusters. Later, we conducted a stability evaluation to identify which suitable clusters could be aggregated into a single dimension.

Conditional Inference Trees (CIT) [67]: Finally, we use tree-structured regression models to explore the relationship between normalized citations and rigor and relevance criteria in the proposed dataset. The method is based on a unified framework for permutation tests proposed by Strasser & Weber [91]. This statistics-based approach uses non-parametric tests (Chi-squared, or $\chi^2$) to test the association between the candidate splitting criteria and the observed value of the dependent variable.

CITs are particularly useful to investigate ordinal variables gathered from subjective human interpretation, such as rigor and relevance scores [67]. Besides a meaningful tool for hypothesis testing, it also provides additional features to analyze and interpret the results. The criterion for testing the hypothesis is based on multiplicity adjusted p-values, thus the stop criterion is maximized (e.g., with a stop criterion = 0.95 the p-value must be smaller than 0.05 in order to split this node). This process also ensures that the right sized tree is grown, requiring no form of pruning or cross-validation.

The CIT model was built using the R environment for statistical computing and the party package, both available from The Comprehensive R Archive Network (CRAN, http://CRAN.R-project.org/).

2.4 Results

A total of 12 SLRs adopting the scoring rubrics to assess reporting quality of the primary studies have been selected, providing 718 instances of primary studies. Both primary and secondary studies identified were inspected to explore two particular contributions of this study.
Towards understanding the relation between citations and research quality in software engineering studies

C1 Overview of the scoring rubrics’ use

The first objective of this exploratory study was to review how SLRs use the scoring rubrics [70] to assess their included primary studies. We explore the use of the rubric focusing on two research questions (Table 2.5):

Table 2.5: Overview of how the Ivarsson’s and Gorschek [70] scoring rubrics have been used

<table>
<thead>
<tr>
<th>Question</th>
<th>Aspect</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RQ1. How was it applied?</strong></td>
<td>As in the rubrics</td>
<td>Refers Ivarsson and Gorschek [70] rubric &quot;as is&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[S10, S13, S14, S15, S16]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Details the scoring rules [S3, S11]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discuss application issues [S2]</td>
</tr>
<tr>
<td></td>
<td>Interpretation of the scores</td>
<td>Maps a checklist [52] to rigor and evidence levels to relevance scores [S6, S8]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Builds objective rules to assess each aspect [S1, S9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Two independent reviewers [S1]</td>
</tr>
<tr>
<td><strong>RQ2. Used as/for...</strong></td>
<td>Quality Assessment</td>
<td>Detailed assessment [S1, S3, S9, S10, S11, S13, S14, S15, S16]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not explicit [S2]</td>
</tr>
<tr>
<td></td>
<td>Objective/Results</td>
<td>Research Question [S3, S9, S11, S16]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discussion of Results [S1, S2, S10, S13, S14]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implication of Findings [S6, S8]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Study Limitation [S15]</td>
</tr>
</tbody>
</table>

*RQ1. How the rubrics were applied:* Most SLRs (67%) used the scoring rubrics as proposed, usually referencing the original work [70]. Two of them (S3, S11) also presented a detailed interpretation of the scoring rules. S2 [46] discussed some issues of applying the rubrics when insufficient information regarding the study is provided in the assessed paper. Four studies used the authors’ interpretation of the scoring rubrics: (i) S1 and S9 improved the rubrics by proposing objective rules to assess the papers; (ii) S6 and S8 used another instrument for quality assessment and then mapped the outcomes to rigor and relevance scores. Finally, S1 [43] discusses methodological issues on conducting the assessment activity with two independent reviewers and then using the Kappa statistic for evaluating the agreement.

*RQ2. The rubrics were used as/for:* The majority of papers (83%) used the scoring rubrics to its proposed objective - to assess the reporting quality of primary studies. Four studies (S3, S9, S11, and S16) formulated a specific
2.4 Results

research question to investigate the reporting quality of studies using rigor and relevance. Five others (S1, S2, S10, S13, and S14) discussed the resulting evidence according to the quality score of the papers. S6 and S8 also discussed the findings, however relating the assessed scores to the implications of the primary studies to research and practice. Finally, S15 discusses the quality of primary studies as a limitation of the SLR process.

It is important to highlight the application of the rubrics in studies S6 and S8, as they differ largely from the remaining papers. Both studies used a subset of Dybå and Dingsøyr’s [52] checklist items to assess the quality of primary studies. We mapped the three rigor aspects to the checklist items, as presented in Table 2.6.

Table 2.6: Dybå and Dingsøyr’s [52] checklist alignment to Ivarsson and Gorschek’s [70] scoring rubrics

<table>
<thead>
<tr>
<th>Rigor aspects [70]</th>
<th>Checklist items [52]</th>
</tr>
</thead>
<tbody>
<tr>
<td>study context (C)</td>
<td>Q1: Is there a rationale for why the study was undertaken?</td>
</tr>
<tr>
<td>study design (SD)</td>
<td>Q2: Is there an adequate description of the context (industry, laboratory setting, products used, etc.) in which the research was carried out?</td>
</tr>
<tr>
<td>validity threats (V)</td>
<td>Q3: Is there a justification and description for the research design?</td>
</tr>
<tr>
<td></td>
<td>Q4: Does the study provide description and justification of the data analysis approaches?</td>
</tr>
<tr>
<td></td>
<td>Q5: Is there a clear statement of findings and has sufficient data been presented to support them?</td>
</tr>
<tr>
<td></td>
<td>Q6: Did the authors critically examine their own role, potential bias and influence during the formulation of research questions and evaluation?</td>
</tr>
<tr>
<td></td>
<td>Q7: Do the authors discuss the credibility and limitations of their findings explicitly?</td>
</tr>
</tbody>
</table>

Moreover, S6 and S8 used a 6-level scale to assess relevance according to the type of evidence provided: 1) no evidence; 2) demonstration or toy example; 3) expert opinion or observation; 4) academic study; 5) industrial study; and 6) industrial evidence. This evidence levels can be related to the relevance dimension in Ivarsson and Gorschek’s [70] rubrics. We, therefore, normalized the evidence levels to the same scale as the relevance dimension (from 0 to 4).
Towards understanding the relation between citations and research quality in software engineering studies

C2 Analysis of the relationship between scientific impact and reporting quality

We conducted a progressive approach to explore the possible relationship between normalized citations and rigor and relevance scores, as follows: 1) a preliminary analysis of the correlation between dependent and independent variables; 2) the identification of underlying structures, i.e., groups of strongly correlated variables; and 3) statistical representation of the relationship based on the observed data.

A preliminary visual analysis is given by building a boxplot of the distribution of the dependent variable on the dataset, as illustrated in Figure 2.3a. The majority of instances (circa 91%) are below 20 citations per year, and 64% have an impact of 5 or fewer citations per year. In addition, the boxplot presents a series of unusual highly cited papers (i.e., outliers).

Further, we conducted a correlation analysis computing Spearman’s $\rho$ coefficient to determine the strength of the relationship between the dependent and each of the independent variables separately. The observed correlation value for rigor is 0.263 denoting a weak ($0.2 < \rho < 0.4$) and positive correlation; and -
2.4 Results

0.017 for relevance, implying a very weak \((0 < \rho < 0.2)\) and negative correlation. Rigor and relevance also show a weak negative relation \((\rho = -0.034)\).

By clustering the variables (Figure 2.3b), it becomes clear that rigor, relevance and normalized citations are fairly orthogonal variables. Rigor is the variable that is more closely related to normalized citations. The paper characteristics are clustered together, suggesting that the length in pages is related to the type of publication (e.g., conferences are often subject to page limitations).

Hierarchical model

The tree model is a hierarchical structure representing the partition of the dependent variable according to splits of the independent variables. Internal nodes are illustrated as circles, expressing the splitting criteria according to an obtained function of association with independent variables (e.g., rigor > 1.5). The terminal nodes (i.e., rectangles) represent the distribution of the dependent variable (normalized citations) according to each split. The size of the tree is automatically determined by the maximum number of internal nodes + 1 to reach the furthermost terminal node. Figure 2.4 illustrates the conditional inference tree model obtained from the data analysis.

The resulting tree contains two internal and three terminal nodes, including all 718 papers. The first node split the instances with a rigor score greater than 2; the remaining instances are split again by node two, on rigor greater than 0 (but still less 2). The left branch of the tree represents the data with low rigor scores (i.e., 2 or less out of 3) while the right side shows papers with higher rigor ratings. One can notice that the relevance criterion has no importance to split the dataset, as no internal node uses the variable as a splitting criterion. Table 2.7 details the terminal nodes, their related splitting criteria, along with characteristics of the papers contained in the node, i.e. median impact value, median paper length (in pages) and distribution of paper types.

Table 2.7: Description of the papers contained in the three terminal nodes.
Towards understanding the relation between citations and research quality in software engineering studies

Figure 2.4: Conditional inference tree describing the relationship between rigor and relevance criteria and the normalized citations.

The terminal nodes with higher rigor scores show a higher median number of citations per year. There is also growth on the length (in pages) and the percentage of journal papers in the nodes with higher rigor scores. On the opposite direction, conference papers are less frequent in nodes 4 and 5. Paper length and paper type characteristics are likely correlated since journal papers are often longer than conference papers.

The resulting tree model and the splitting subsets show that studies with high rigor (i.e., on the right side of the tree) often achieve greater citation impact compared to papers with lower rigor scores. Thus, the null hypotheses regarding the rigor aspect is rejected, indicating that there is a significant relation between that criterion and the dependent variable (with a p-value of at least 0.001).
relevance aspect, however, has no significant relationship (at $\alpha = 0.05$) with the normalized number of citations.

**Distribution of Citation Counts**

In need to better understand the distribution of the citations according to each terminal node, we refine the earlier cluster analysis, also omitting the outliers (as illustrated in Figure 2.5). Looking in more detail into this helps us understand how the impact variable is distributed on each splitting subset.

The density plot shown at the bottom of Figure 2.5 presents the distribution curve for each node. It is clear that node 5 has a wider distribution, assuming that papers present a higher rigor score can achieve a higher number of citations. The distribution of citations on Nodes 3 and 4 is right-skewed, thus showing that few studies on these nodes achieve a high citation count.

**Sensitivity Analysis**

Further, we conducted a sensitivity analysis to test the robustness of the tree model, and increase our understanding of the relationships between the dependent and independent variables. The analysis was conducted by recalculating the outcomes under the alternative assumptions that the different studies can influence the results by introducing a significant amount of biased data (e.g., low rigor papers).

We recalculate the outcomes by removing one factor (i.e., the instances related to a particular study) at a time and comparing the produced output with the conditional inference tree presented in Figure 2.4. Most of the resulting alternative trees are visually similar to our model, presenting three terminal nodes and an increase to normalized citations related to higher rigor scores. Two alternative models (i.e., removing the instances from S1 and S13) presented only one splitting node regarding rigor, still resulting in the same pattern on the normalized citations.

Three alternative trees (i.e., removing the factors from S6, S11, and S14) resulted in models that largely differs from our original model, as shown in Appendix 2.9.1. All of them include an internal node for the relevance variable (with p-values of 0.034, 0.031 and 0.019, respectively), further adding splitting criteria to a subset of the rigor variable. We further grew another alternative model by removing the factors from S6, S11, and S14; to verify the impact of these instances to the model. Surprisingly, this last model is very similar to the original CIT, as shown in Figure 2.6.
Towards understanding the relation between citations and research quality in software engineering studies

Figure 2.5: Boxplot and Density plot for the impact (i.e., normalized citations per year) according to the splitting nodes, omitting outliers.

At large, the p-values and splitting criteria of the alternative models slightly differs from the original model (e.g., ri > 1.5 instead of ri > 2). The outcomes from the alternative models do not invalidate our process but raise some additional discussion on the features that could influence the results. Small divergences among the models are expected due to the differences in the number of instances, the reporting quality scores and the subjective nature of the quality assessment. Overall the sensitivity analysis shows that our model is robust.
2.5 Discussion and Implications

This exploratory study provides two distinct contributions:

C1 Identifying methodological issues of the application of the Ivarsson and Gorschek [70] scoring rubrics

The main purpose of the rubrics is to assess the quality of primary studies during the conduction of SLRs. However, often researchers also address the research questions and implications of the study according to rigor and relevance scores, as demonstrated in Section C1. Therefore, it is valuable to investigate if different quality assessment instruments (such as [39, 52]) have been used in a similar manner. This could imply a particular need to relate the assessed quality of
Towards understanding the relation between citations and research quality in software engineering studies

primary studies with the evidence collected by the SLRs, as well as to propose new guidelines or update the existing ones for a comprehensive evaluation of reporting quality.

SLRs adopting the rubrics often describe the researchers’ experience and methodological issues. However, no formal evaluation of the proposed instrument was done. Despite referring to the rubrics [70], some studies further detailed the scoring rules (S3 and S11) or even proposed some objective support (S1 and S9) to interpret the rigor and relevance criteria, thus suggesting issues related to its subjectiveness. We highlight the need for further evaluation of the scoring rubrics, investigating such issues and validating possible solutions. In particular, several researchers should score the same papers and evaluate whether they come to the same conclusions across a large set of studies. This gives confidence in the objectivity of the rubrics. Given that the rubrics are on a relatively abstract level, there may be a need to complement them with additional checklists as has been done by S1 and S9. Furthermore, S3 and S11 have seen a need to refine the rubrics. This may also affect the ability to assess the papers objectively, thus reducing biases in the assessment.

C2 Investigating the relationship between rigor and relevance

The results of the statistical analysis show that the relation between only rigor and scientific impact is positive. Evidence obtained with the conditional inference tree model suggests a positive relation between rigor and the normalized citations per year. Despite this, we cannot infer that this relationship represents causality (i.e., increasing rigor causes an improvement on citation counting), as rigor is likely not the only criterion to impact the number of citations.

The build model also shows a lack of a significant relation between relevance and normalized citations. This result does not mean that relevance is negatively impacting the citations, but otherwise suggests that SE studies with high relevance were not particularly acknowledged by the research community with a higher number of citations. A plausible reason for this lack lies in that researchers are not aware of the relevant demands from industry [63,73]. Therefore, it is important to encourage researchers to conduct and to evaluate studies according to the potential impact equally on academia and practice.

Although several guides foster the conduct of realistic studies in SE (e.g. [83, 89,90]), this is particularly challenging as it often requires a representative of the real usage, the involvement of practitioners, and an industrial scale. Moreover, research methods that produce results potentially more relevant usually come
up with a lack of control, likely implying that achieving one criterion well may have a negative effect on another [77]. Our results showed some studies scoring high on both rigor and relevance, demonstrating that it is possible to ally rigor and realism.

Industrial practitioners ought to be interested in studies that have high relevance. Despite that, existing studies suggest that practitioners rarely access academic literature [56,63,85,102]. Consequently, to influence and value of high relevance research for practitioners, there is a need to communicate the findings to practitioners in a different way. Potential ways could be blogs, tweeting the latest research findings and linking to the sources. This would also allow measuring the impact of academic work through the feedback that could be obtained from practice through the networks.

Ultimately, the results suggest a need for better identifying and assessing the industrial impact of SE research. It is likely that this need also occurs in different applied research fields, or even across other fields of science. Cross-domain replication studies are important to test and compare our results, thus refining general findings from the particular SE-related results. For conducting such studies, proper instruments to assess the quality of studies are necessary. Ivarsson & Gorschek’s [70] rubrics are proposed specifically for the SE domain, and its suitability to different fields have not been investigated.

A multidisciplinary approach to research quality standards is proposed and further evaluated by Mårtensson et al. [77]. This conceptual model may be the basis to identify a comprehensive set of quality criteria to assess research practice fairly. It is necessary to break its high-level concepts into finer criteria that represent actions ensuring research quality. As part of our future work, we aim to investigate the relevance of such quality criteria, in terms of its attributes, to different types of research methods employed in empirical software engineering; our goal is to support the ESE community in making informed decisions in the design of research across different SE sub-areas and groups of research.

2.6 Validity Threats

A series of issues may influence the results of this exploratory study, such as the researchers who performed the study, the observed dataset and the measures of reporting quality and scientific impact. In the following, we consider the threats to validity, as a way to discuss the acceptance and accuracy of our findings.
Towards understanding the relation between citations and research quality in software engineering studies

Internal validity

The scoring rubrics were not yet evaluated by the community, though they were used to assess the reporting quality of primary studies in several SLRs and SMs. We rely on the community’s use to vouch for the acceptance of the rubrics. Given that the rubrics were not evaluated, different researchers have used them in different ways. The actual interpretation of different researchers may not be aligned. Hence, a score in one investigation may not mean exactly the same as the same score in another study. In Section C1 we presented how the reviewers adopted the scoring rubrics when assessing studies.

Moreover, in our exploratory work, we identified three candidate studies that assessed the rigor and relevance criteria differently from proposed by the scoring rubrics. S1 considered two of the rigor aspects (i.e., context and research method) as selection criteria, as S6 and S8 adopted different assessment instruments [53], mapping the scores to the rigor and relevance aspects of the scoring rubrics. We also identified three relevant studies (S13, S14, and S16) originated from master and Ph.D. theses, i.e. they were not peer-reviewed. Despite this, those papers presented assessment scores and paper characteristics similar to the remaining, peer-reviewed studies. The difference in the assessment of each of these studies might have an impact on our results.

To address such potential bias, we conducted a sensitivity analysis of the statistical model (Section C2). Three alternative assumptions produced by removing one source of data (S6, S11, and S14) result in a model non-compliant with the hierarchical model. Those alternative models show extra splitting nodes related to the relevance criteria. However, most of the recalculations (9 out of 12 alternatives) produced similar results to our original tree model.

External validity

The data was gathered from a set of SLRs and SMs using the scoring rubrics for quality assessment of included papers, which introduces a potential bias. Though in order to (a) assess both rigor and relevance, and (b) achieve a high counter of primary studies, we based our analysis on the scoring rubrics by Ivarsson and Gorschek [70]. That is, all papers included started from the same description on how to assess rigor as well as relevance.

We do not claim that our results are representative of all software engineering literature, or to be generalizable to other fields or contexts. Though, it provided insights from a wide range of topics in SE (including economics and professional
practice) across several SLRs. Also, the total number of primary studies was high (over 700).

Further studies using the rubric will get published in the future, which may influence the results. Though, with our sensitivity analysis (Section C2) we have shown that the results appear to be robust when removing or adding studies to the set. Hence, this provides confidence in the findings of our study.

**Construct validity**

Citations as a measure of scientific impact have been criticized. There are multiple purposes for citing a study, such as assumptive citations (referring to general knowledge), affirmational (confirming existing findings), contrastive (contrasting findings with existing work) and methodological (building and using method guidelines from existing work) [48]. Though, the judgment of citation types is subjective and often inconsistent among reviewers [84]. In this Chapter, we did not make the distinction, though it may be fair to assume that an assumptive citation may not rely as much on high rigor and relevance in comparison to a methodological paper.

Beyond that, mature research papers have a head start over the novel studies as their citation counts likely increase as the years pass [97]. The citations evolution is not linear though, and one can argue that our scenario is unfair to the most recent papers (i.e., published in 2013). We believe that a delay of two years for the data collection does not impact the results significantly, on the contrary, is likely to hide the peculiarities of novel research (e.g., papers frequently cited shortly after their publication). Moreover, this scenario provides a representative sample that reflects the actual state of the art, in which both novel and well-established research are represented.

The scoring rubrics rely on a subjective evaluation of rigor and relevance by the reviewers. Such subjectivity could result in divergent scores depending on the reviewer’s previous experience and knowledge. In our study, we identified three duplicated primary studies assessed by different researchers in different SLRs. All of those present small divergences in the rigor or relevance scores (e.g., medium (0.5) rather than strong (1) description for the study design). One can reason that such divergences in the reviews are due to subjective interpretation of the scoring rubrics.

Moreover, rigor and relevance are not the only criteria related to the quality in the reporting of research. Additional confounding factors could be strongly related to the number of citations. Paper characteristics (i.e., length of the paper, type of publication) and research factors (i.e., research method, context,
Towards understanding the relation between citations and research quality in software engineering studies

industrial applicability) showed a relation with the number of citations. We briefly addressed this validity issue in Section C2. A dendrogram demonstrated that two investigated papers characteristics (i.e., length in pages and type of publication) are related.

The impact of such and other confounding factors on the dependent variable were not further investigated, as not all of them are feasible to identify and extract. Furthermore, the rigor and relevance criteria are focused on what is reported and do not completely cover all relevant actions to be taken to, for example, evaluate a controlled experiment or case study in depth. Though, as a consequence, the rubric can be applied to diverse study types, hence including more than 700 papers, and allowing to focus on rigor and relevance at the same time.

Conclusion validity

During the conduct of the experiments we mostly used a single researcher to fetch the papers, collect the data, built the tree model and analyze the results. Most importantly, as the same author drew the conclusions from the gathered data, there is a risk related to the interpretation of the findings. We tried to mitigate this validation threat by discussing the preliminary results at length with the second author, and further reasoning our conclusions with the third author.

2.7 Conclusions

In this Chapter, we presented an empirical study to explore the relationship between scientific impact and quality in the reporting of research. Our investigation was conducted on the findings of systematic literature reviews and mapping studies that use rigor and relevance criteria for quality assessment of primary studies, as proposed by Ivarsson & Gorschek [70]. Based on two distinct contributions, the findings of this study are:

C1. We identified 16 SLRs using the scoring rubrics to assess the reporting quality or classify primary studies; wherein 12 provide detailed information on the application (i.e., the scores for each assessed study). We analyzed the selected SLRs assessing how the rubrics were applied and for which purpose. Our findings suggest that the scoring rubric could benefit from empirical evaluation and further refinement. The use of complemen-
tary instruments also suggests that the rubrics are an early-stage quality evaluation, requiring more specific assessment rules.

C2. After that, we provided a statistical analysis of the relationship between rigor and relevance scores and the normalized citations. Evidence implies a contribution for scientific impact with increasing rigor of the studies (i.e., how the research is conducted and reported). Although we cannot elucidate this relationship at this time, the results support the raising of a question regarding the importance and worth of research showing potential impact on academia and industry, i.e. factors addressed by the relevance score. This study provides a foundation to discuss and reflect on the current findings. For example, we may reflect on whether we should strive for a stronger relation between rigor and relevance. In case we come to the conclusion that this relation is, in fact, important, then further steps need to be taken to strengthen it. Examples could be to reflect on selecting primary studies according to the relevance criterion when conducting systematic reviews and mapping studies.

As future work, we encourage the replication of this study across different disciplines in order to better understand the link between research quality and scientific impact. We believe that further development and evaluation of multidisciplinary instruments to assess research quality are required. Overall this line of research has the potential to lead to discussions and consensus building on how to fairly and accurately reward high quality in research practice.

Acknowledgements The authors would like to acknowledge Prof. Tony Gorschek, Prof. Niklas Lavesson and Dr. Michael Unterkalmsteiner who contributed to the discussions that led to this study. We also would like to thank Joakim Pernstål, Matthias Galster and Vahid Garousi who kindly shared the data of the included studies.

2.8 References


Towards understanding the relation between citations and research quality in software engineering studies


2.8 References


Towards understanding the relation between citations and research quality in software engineering studies


Towards understanding the relation between citations and research quality in software engineering studies


[87] Saripalli Phani Shashank and David Hem Paul Darse. Finding common denominators for agile software development: a systematic literature review.


Towards understanding the relation between citations and research quality in software engineering studies

2.9 Appendix

2.9.1 Sensitivity Analysis of the CIT

Visual representations of the alternative models grown for sensitivity analysis, as detailed in Section C2 of Chapter 2:
2.9 Appendix
Towards understanding the relation between citations and research quality in software engineering studies
Chapter 3

CERSE - Catalog for empirical research in software engineering: a systematic mapping study

This chapter is based on the following paper:


3.1 Introduction

Methodological guidance for research practice is more often provided in the form of procedures or guidelines i.e., a step-by-step description of the particular research process and its underlying actions. Assessment instruments are sometimes used to ensure that the process is followed according to particular evaluation criteria.

The challenge of selecting a suitable research method and furthermore to identify appropriate guidelines to apply such method is not new to Empirical
Software Engineering (ESE) [358,421]. Researchers such as Shaw [115] highlight the lack of guidance in designing research and organizing the reporting of ESE studies.

More recently, Wohlin [119] recommended a set of commitments that researchers should be aware of when conducting their studies. One of these commitments is mainly related to using and following available guidelines when conducting research. Ultimately, this commitment raises another issue relating to how to identify the appropriate guidelines and other related methodological support.

In the field of medicine, there has been an initiative [117] to address such issue. The EQUATOR Network comprises a comprehensive catalog of guidelines for health research. We believe that a similar initiative in SE will raise researchers’ awareness regarding the wider range of available guidelines for the methods they are using [119,421], hence supporting them to make informed decisions of how to design their research.

Thus, this mapping study contributes to the body of knowledge in SE by proposing a catalog of existing guidelines, assessment instruments, and knowledge organization systems (see a glossary of terms in the Appendix 3.8.1) supporting different kinds of empirical methods. We further make the catalog available to researchers by means of the CERSE Web tool\(^1\), which also provides searching and filtering functionalities. The Web tool implements a step-by-step strategy to identify and select the guidelines more suitable for the particular research. We intend to update the catalog periodically, so making the more recent methodological support also available in our tool.

The remainder of this Chapter is organized as follows: Section 3.2 summarizes similar, and related work; Section 3.3 describes the methodology we employed at searching and selecting the relevant literature for the catalog. In Section 3.4 we present the findings according to the research methods, data collection and data analysis methods related. Later, Section 3.5 describes a process for using the catalog to select a set of proper guidelines, and discuss the implications for research practice. We conclude our study in section 3.6 providing some suggestion for further work.

\(^1\)Available at: http://www.cerse.org
3.2 Related Work

3.2.1 Research Methods Selection Procedures

Early stages of a research design require from researchers many decision-making steps (regarding hypothesis formulation, context definition and research methods employed). They are not trivial choices, as they impact the research process and are likely to shape research results. On the selection of a Research Method, it is essential to recognize the characteristics, advantages, and limitations of each methodology.

Creswell’s work [103] is a well-established and recognized literature addressing such issue. The book addresses the whole research process and its underlying steps, further proposing a framework for research design in human and social sciences. Regarding the selection procedure, it provides support on choosing from quantitative, qualitative, and mixed methods designs and methods. In software engineering, Shull et al. [116] elicited practical and theoretical knowledge for conducting, reporting and using empirical research methods.

Additionally, specific literature addresses the research method selection issue in software engineering. Eastbrook et al. [104] describe five empirical methods available (i.e., experiment, case study, survey, ethnography and action research), identifying the types of questions best addressed by each one. Kitchenham, Linkman and Law [285] proposed DESMET, a methodology for evaluating software tools and methods that employ different research approaches combined. The guidelines support researchers at identifying the appropriated research units and methods to use.

Recently, Wohlin and Aurum [449] propose a framework to support researchers to decide upon which methodologies to employ at strategical, tactical and operational levels. Data collection and data analysis techniques are listed along with more comprehensive research units, such as case study and action research. Relevant literature to each method is also presented, some of which include actual guidelines to conduct the process.

3.2.2 Aggregations, Classifications and Secondary Studies on Methods in Software Engineering

Studies that extensively investigate a research field in search for methodological guidance are comparable and related to this work. Their importance lies in providing researchers with an opportunity to consider a comprehensive set of available guidelines for their research.
Secondary and tertiary studies aggregating guidelines and supporting literature for research emerged across different fields, such as healthcare [105,111,117] and economy [114]. These studies provide an overview of the state-of-art from a method research’s perspective; however, they often focus on a specific context e.g., decision-analytic modeling [105] and enterprise development [114].

As previously mentioned, a particular example of such aggregation is the EQUATOR initiative [117]. It provides a searchable database for researchers interested in conducting different studies. The catalog also presents extensions to the main guidelines, providing additional support for its particular phases or contexts e.g., controlled trials, longitudinal observations.

Aggregations resulting from secondary studies on methods are also not unknown in software engineering. Kitchenham & Brereton [106] conducted a systematic literature review (SLR) to identify published experiences of performing such reviews in the SE context. The resulting 63 studies are analyzed in relation to their contributions to each particular task of the SLR process. This work differs from ours by focusing only on the guidelines’ extensions obtained through experience reports for a specific research method i.e., SLR.

Furthermore, Borges et al. [148] aggregate and classify supporting mechanisms adopted by empirical studies published in major ESE scientific venues (i.e., EASE, ESEM, and ESEJ). A total of 412 mechanisms were identified since 1997, addressing experiment (42%), case study (24%), survey (7%) and other 11 research methods. Although this study has a similar aim to ours, its results differ from ours in a number of ways:

i) we aimed to specifically investigate the guidelines published in the SE domain, instead of assessing the specific individual studies that have employed such guidelines;

ii) we covered a broader set of venues;

iii) we employed snowballing strategies to identify guidelines from other disciplines cited in ESE context (e.g., Information Systems);

iv) we also extracted the additional information we believe is vital for researchers when selecting the appropriate guidelines. Such information include: type and objective of the instrument, process phases addressed, the maturity of the instrument, and relation with other research methods; and

v) we aggregated a catalog of guidelines and supporting instruments for empirical research, which is available online via a Web tool¹.
3.3 Method

Our review protocol was developed according to the guidelines for conducting systematic literature reviews (SLRs) and mapping studies (MSs) in software engineering by Kitchenham et al. [109,294]. Those guidelines have been extensively used in SE secondary studies, and further validated through a series of papers (e.g., [128,134,155]).

All the researchers involved in this study took part in the development of the protocol. We conducted several meetings to discuss the activities of the MS process, such as specifying the research question, allocating tasks to each researcher, planning the search and selection process, data extraction and synthesis. We also ensure that all changes during the execution of this mapping study and their implications were discussed among the three authors and further updated in the protocol.

3.3.1 Research Question

The objective of this MS is to aggregate and report evidence about existing literature on guidelines, assessment instruments and knowledge organization systems (see a glossary of terms in the Appendix 3.8.1) for conducting and evaluating empirical studies in SE. To achieve the mapping study’s main objective we formulated the following research question:

RQ. What are the available guidelines, assessment instruments and knowledge organization systems for empirical research in software engineering?

Note that, it is important to categorize and group the supporting literature from multiple perspectives (namely research methods, phases of the research process, and maturity). The categorization is intended to make explicit the coverage of the research field according to the different perspectives.

In general, SLRs and MS report evidence gathered from primary studies, i.e. studies that report the results of empirical research by means of, for example, case studies, surveys and experiments. However, within the context of this study we are using the term “included studies” instead, because our results (i.e. guidelines and other supporting literature) are publications that provide methodological support to conduct the so-called primary studies.
3.3.2 Search and Selection Process

First, we carried out a database search, as suggested by Petersen, Vakkalanka & Kuzniarz [113]. However, the results from this automated search, after using numerous different combinations of search strings, did not retrieve a set of already known papers. Therefore, we chose to carry out a manual search including a backward snowballing strategy.

We adopted a combination of search and selection strategies [360,447,458]. This mixed approach is intended to identify as much relevant evidence as possible, to ensure the repeatability, and to evaluate the search and selection process. An overview of the search and selection process is presented in Figure 3.1, where A1, A2 and A3 represent the first, second and third authors, respectively.

**Quasi-gold standard.** Prior to conducting the search process, we identify a collection of already-known papers. This quasi-gold standard comprises guidelines to the main research methods in SE domain: action research [315], case study [288,391], experiment [365,456], interviews [398], observation [309], survey [108], systematic literature review [294] and systematic mapping [361].

**Piloting a database search.** Based on the quasi-gold standard, we derived a search string that we used to pilot 4 major search engines, as proposed by Kitchenham et al. [109]: 1) ACM Digital Library 2) EI Compendex 3) IEEE Explore 4) Scopus (which indexes the references by a series of publishers e.g., Springer and Wiley).

The results were not satisfactory due to the high number of false positives, given that not only guidelines but also studies following them were returned. We later produced alternative versions of the search string, but the results did not include all the papers within the quasi-gold standard. Moreover, the number of resulting papers was too high for a reasonable selection process (i.e., more than 100,000 papers).

**Manual search.** Our first search attempt i.e., a standard database search, provided a high amount of true negatives and false positives. Most of the search results comprised papers that do not present guidelines or other methodological support for conducting empirical research. Hence, the more effective alternative was to start with a manual search on the main fora where many papers with a methodological focus are published, namely:

**EASE** International Conference on Evaluation and Assessment in Software Engineering (1997-2014)

3.3 Method


**IWSM-Mensura** International Workshop on Software Measurement (1991-2014)

**METRICS** International Software Metrics Symposium (1993-2005)


Figure 3.1: Overview of the search and selection process. On each step (i.e., manual search and snowballing iterations) we identify a list of candidate studies for selection. Those candidates already excluded any irrelevant paper (i.e., according to exclusion criteria E2, E3, and E4). After each study selection stage, the included papers identified are aggregated into a joint list.
Through the snowballing process, additional venues that are not part of the starting set provided relevant papers. Some of these venues are Information and Software Technology Journal (21 included papers), ACM SIGSOFT Software Engineering Notes (20 included papers) and IEEE Transactions on Software Engineering (12 included papers).

Selection criteria. After identifying the main fora, we specified a set of selection criteria. These inclusion and exclusion criteria are meant to classify the candidate studies, and also to select those that are suitable to be included in the catalog i.e., answer the research question in Section 3.3.1.

- **Inclusion criteria:**
  - I1 methodological articles for conducting empirical research;
  - I2 checklists and instruments to assess the quality of empirical research;
  - I3 ontologies, taxonomies, and artifacts to classify and characterize empirical research; and
  - I4 articles reporting lessons learned and/or recommendations to improve the research process.

- **Exclusion criteria:**
  - E1 nonpeer-reviewed journals, conferences, workshops, and books (i.e., doctoral symposium papers, posters, workshops, and keynotes);
  - E2 papers outside to software engineering and related domains (e.g., information technology, computer science, and systems engineering);
  - E3 papers not written in English; and
  - E4 duplicated and not available papers.

Validating the selection criteria. We further employed a “think-aloud” for selection process on a random sample of the candidate papers to develop a shared understanding between the reviewers. The first author (from hereon A1) performed the selection documenting the reasons to include or exclude a paper while the second author (i.e., A2) acted as an observer. The reasoning was later discussed with the third author (i.e., A3) to help improve and minimize any ambiguities found in the selection criteria.

Piloting the selection. Further, another selected subset of papers was assessed independently by each author as a pilot. Our subset consisted only of papers published in the EASE 2014 and ESEM 2013 conferences, totaling 114
3.3 Method

papers. Details about the choice were documented this time, but the disagreements were analyzed using different scenarios for study selection, as proposed by Petersen & Ali [360]. An inter-rater agreement level is calculated based on the outcomes, and as it was considered good (observed kappa = 0.693), the actual study selection started.

Study selection process. In this step, each paper was simultaneously assessed based on titles and abstracts by two authors i.e., A1 assessed all identified papers, while the A2 and A3 each assessed half of the studies. We applied a set of rules for deciding on whether an article was to be included or excluded, as shown in Table 3.1. Our rules were inclusive, meaning that minor disagreements (i.e., include-uncertain) were solved by including the paper. Further, only papers excluded at least by two authors were removed without any additional assessment. A single vote on excluding a paper (or two uncertainties) led to a disagreement solved by the other author, who had not participated in the selection of the paper.

Table 3.1: Decision rules to study selection. All papers were independently classified as included (I), excluded (E) or uncertain (U). By comparing the assessment of two reviews, we arrive at a decision whether include or exclude the paper. The disagreements (D) were solved by an additional decision vote whether included or excluded.

<table>
<thead>
<tr>
<th>1st reviewer</th>
<th>I</th>
<th>U</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd reviewer</td>
<td>I</td>
<td>I</td>
<td>D</td>
</tr>
<tr>
<td>U</td>
<td>I</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>E</td>
<td>D</td>
<td>D</td>
<td>E</td>
</tr>
</tbody>
</table>

Snowballing search 1. The results of the study selection process established a collection of potentially relevant sources for backward snowballing, according to Jalali & Wohlin [249]. The source for snowballing was substantial (150 included papers) and clearly fulfilled the criteria for heterogeneity of the starting set (i.e., different authors, publication venues and covered years).

Our snowballing approach consisted on reading specific sections of the papers (i.e., background, related work, and methods) likely to provide methodological references. Then, we collected the candidate references, excluded the
CERSE - Catalog for empirical research in software engineering: a systematic mapping study

duplicates (among the ones identified in the same snowballing iteration plus the ones already included by previous iterations) and papers that otherwise were not available to retrieve (see exclusion criteria E4). The resulting 287 references were aggregated into a candidates list and assessed according to the study selection process described above. At the end of this process, 155 additional papers were included.

Snowballing search 2. A second snowballing iteration followed the same process as snowballing search 1. Papers included in the previous iteration were our starting set, from which we retrieved 185 candidates. After selection, 36 new papers were added to our included data set.

Analysis of included papers. Once the snowballing search and selection process was finished, we compiled a list of the included papers and analyzed their particular characteristics (e.g., year and venue of publication). Although our starting seed contained just 6 fora (5 conferences plus 1 journal), the snowballing strategy extended our reach to 75 additional venues (59 after the first iteration step plus 16 after the second iteration).

Search and selection performance. Finally, we compared the proportion of resulting candidates (diminishing returns henceforth) and quasi-gold standard (quasi-sensitivity henceforth) in the retrieved references to an optimum threshold of less than 20% and greater than 80%, respectively, as described by Zhang and Babar [458]. The threshold was not reached by the manual search, or by the first snowballing iteration (Table 3.2), thus we performed the snowballing process a second time.

Table 3.2: Analysis of the performance of search and selection process. Diminishing returns is defined as the proportion of candidate studies in the retrieved references, whereas quasi-sensitivity is calculated by comparing the quasi-gold standard to the included papers.

<table>
<thead>
<tr>
<th>Iteration Step</th>
<th>Manual</th>
<th>Snowballing 1</th>
<th>Snowballing 2</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved references</td>
<td>2114</td>
<td>744</td>
<td>997</td>
<td>3855</td>
</tr>
<tr>
<td>Candidate studies</td>
<td>2109</td>
<td>287</td>
<td>185</td>
<td>2581</td>
</tr>
<tr>
<td><strong>Diminishing returns</strong></td>
<td><strong>99.8%</strong></td>
<td><strong>38.6%</strong></td>
<td><strong>18.6%</strong></td>
<td><strong>67%</strong></td>
</tr>
<tr>
<td>Already-known papers</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Already-known included</td>
<td>2</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td><strong>Quasi-sensitivity</strong></td>
<td><strong>20%</strong></td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
3.3 Method

The decrease in the number of diminishing returns of the second snowballing iteration implies that we were approaching a critical mass i.e., the number of retrieved references will exponentially grow until it reaches the stage beyond which no further study is selected. A high degree of saturation is desirable, as it reflects the completeness of the search process. However, a similar exponential effort would be needed to identify and select such critical mass, that is likely to provide a negligible number of candidate studies.

The quasi-sensitivity validated the completeness of our search strategy regarding the quasi gold-standard. During the manual search, just 2 out of the 10 already-known papers were included. At the end of the first snowballing iteration, all papers in the quasi gold-standard were identified and included. This does not imply that all the relevant papers within the search universe were included, instead, it suggests that our search performance is acceptable.

3.3.3 Data Extraction

The data extraction stage comprises the identification and gathering of information needed to address the research question. The first author designed the data extraction instrument in accordance with the research objective and the intended categorization. Further, all the three authors discussed design issues that should be considered. After a few refinement iterations, the data extraction instrument was judged as ready for use. This instrument includes records of relevant information from the included studies, as well some additional bibliographic data, as follows:

D1 Paper title;
D2 Authors and affiliation;
D3 Year of Publication;
D4 Paper type (i.e., conference or journal) and publication venue;
D5 Name and summary of the guideline or supporting instrument targeted by the paper;
D6 Research methods to which the instrument can be applied i.e., experimentation, survey, case study;
D7 Process phases covered by the guideline or supporting instrument i.e., planning, execution, analysis and/or reporting;
D8 Aim of the supporting instrument e.g., to conduct, to classify, to evaluate or to replicate research; and
D9 Maturity score i.e., classification of research facets according to Wieringa et al. [445].

D6. Research methods. Aiming to consolidate the knowledge regarding research methods and their relations we investigated the previously identified related work already [148, 358, 449, 456]. However, we found out that different studies presented differing points of view relating to research methods, producing divergences in our classification. Thus, we decided for an initial classification according to [449], updating the classes as needed during the data extraction process. We also validated the new labels in comparison to the existent classifications [104, 421, 449]. Further, we reviewed the entire classification, aggregating research methods by similarities described in the included papers. Further information regarding the research methods and their relations is given in Section 3.4.1.

D8. Aim of the supporting instrument. To better understand how the identified instruments can support the research process, we propose a classification according to their intended objectives, as follows: 1. guidelines for conducting research; 2. assessment instruments to evaluate research; and 3. knowledge organization systems to classify or characterize research. For details of this classification see a glossary of terms in the Appendix 3.8.1.

D9. Maturity score. To determine the contribution of each guideline and instrument to our study, we distinguish between proposed and evaluated solutions (see a definition of these terms in the Appendix 3.8.1). Each included study was assessed using the classification of research facets by Wieringa et al. [445], as recommended in Petersen et al. [113]. The classification provides a means to identify mature research, as well as instruments that need further investigation:

- **Problem investigation:** opinion paper (stated position).
- **Solution design:** philosophical paper (conceptual framework).
- **Solution validation:** validation research (weak empirical study).
- **Solution selection:** experience report (lessons learned).
- **Solution implementation:** solution proposal (proof-of-concept).
- **Implementation evaluation:** evaluation research (strong empirical study).

Pilot extraction. Prior to the data extraction, the three authors performed a pilot data extraction on a random sample of five selected studies. For each paper, all three authors separately extracted the relevant information (see Section
3.3 Method

3.3.3) comprising mainly qualitative data. We were able to assess the degree of agreement related to the categorical data i.e., D6, D7, D8, and D9, as follows:

D6 Regarding the research method there were two cases of partial disagreement, in which A1 classify the paper into two research method categories (e.g., case study and experiment) and one of the second authors related to just one.

D7 A partial agreement is reported in just one case, due to a set of multiple process phases.

D8 No disagreement is reported regarding the aims of the supporting instrument. However, our sample consisted only of guidelines for conducting research.

D9 One disagreement was reported in relation to the maturity score.

Further, we compared our data collection forms and discussed the disagreements, so to reduce any possible data extraction-related validity threats. We consider that the partial disagreements are not a threat to the data collection, as the first author presented the most comprehensive set. The single disagreement of D9 is due to a divergent interpretation from one of the second authors. We scheduled periodic meetings so A1 could discuss further data extraction with the other two authors.

Validation of data extraction. The extraction process was carried out by the first author of this Chapter and reviewed by the second author, as suggested by Kitchenham et al. [109]. After the data extraction, a sample of 10% (i.e., 37) of the included papers was validated independently by the second and third authors. The outcomes of the validation were further discussed by all authors.

The validation identified uncertainties regarding a study’s classification due to the information provided in the paper. These uncertainties are related to the multiple classifications of the included papers regarding (D6) research methods, (D7) process phases, and (D9) aims of the instrument. Several papers can be classified in more than one category as they provide support for a set of research methods (e.g., a case study using observations for data collection), research phases (e.g., guidelines for analysis and report of studies) and aims (e.g., both to conduct and evaluate the research).
3.3.4 Analysis and Classification

The extracted data was further tabulated and arranged so to be used to answer the research question, as follows:

RQ. What are the available guidelines, assessment instruments and knowledge organization systems for empirical research in software engineering? The available guidelines, assessment instruments and knowledge organization systems for empirical research in SE were compiled in a reference list provided in the Appendix 3.8.2. Further, we summarize the findings according to three perspectives:

- **Research method.** An overview of included papers related to each research method is presented and discussed in Section 3.4.1.

- **Phases of the research process.** An overview of the coverage of process phases (i.e., planning, execution, analysis, and reporting) by guidelines and other supporting instruments is given in Section 3.4.2.

- **Evaluation.** More mature instruments are implemented and evaluated in practice rather than proposed and validated through toy examples. In Section 3.4.3 we discuss the instruments that achieve a better classification according to our maturity score.

3.3.5 Validity Threats

Some potential threats that could affect the validity of this mapping study and corresponding mitigation strategies are detailed below. The threats were classified according to Maxwell’s categorization of validity threats [110]:

**Descriptive validity**

To ensure the objectivity of the process, we undertook several quality measures. First, a **review protocol** was iteratively developed and updated during this study. During the planning phase, we conducted a series of meetings to ensure the protocol completeness and understandability. During the execution and analysis phase, additional meetings addressed responses to changing events. After each meeting, the protocol was updated by A1 and further shared with A2 and A3 for review.

Our **search strategy** combined both manual search and snowballing technique. A manual search was performed at some venues with a specific focus on the development and improvement of methodology in SE domain. Further, two
snowballing iterations extended our search beyond those venues, still limited due to the references in the papers included in the previous iteration.

Although we cannot ensure that all the available papers were collected, we employed two actions to validate the search process:

i) before the search activity, we identified a set of already known studies that were further compared to the selected papers; and

ii) the diminishing returns on the search and selection were used as a decision criteria for the snowballing iterations.

Our snowballing strategy does not aim to retrieve all the references from the candidate papers, rather a subset of the references contained in the background, related work and methods sections. We assume that other sections are unlikely to cite any new methodological references. We also piloted the selection criteria through a think-aloud exercise with a sample of the identified papers. Further, the study selection process involved two of the authors, and further disagreements were mediated by the third author.

The data extraction form was piloted by all the authors on a sample of included papers. We compared the extraction form to ensure that consistent information was retrieved from us and discuss any divergences. Similar discussions were carried out during the actual data collection, conducted mainly by A1. The first author has prior experience with the SLR process and theoretical understanding of multiple research methods. However, a reliability threat exists whenever the researcher may be influenced by their prior experiences and the clarity of the papers being extracted. The results of data extraction were further discussed with the other two authors in an attempt to minimize any judgmental error that may have happened while categorizing the research methods, related phases, aims, and maturity score.

The systematic process of secondary studies is intended to allow readers to assess its repeatability. Further replication or extension of this review should be grounded on the details provided in the review protocol and employed guidelines. The steps we undertook to carry out this study are detailed in the methods section 3.3, and the guidelines are provided mainly by Kitchenham & Chartes’ [294].

Interpretive validity

The interpretative validity is related to the objectiveness of data representation aiming to infer conclusions. To interpret the results of this systematic mapping
study, we relied on visual and tabular representations of the data. Moreover, the included studies were classified according to the research methods, process phases, aim, and maturity level. This classification allowed us to group the data and to investigate trends.

Our research question is qualitative in nature, and the categorization perspectives could be addressed by counting the number of included papers in a particular group. Ultimately, we aggregate a comprehensive list of the identified methodological studies into a catalog for empirical research.

Theoretical validity

To minimize the potential theoretical threats related to the uncertainty of collecting suitable data to answer the research question, we proposed a study design through careful deliberation and reasoning. Before starting the systematic mapping study, we detailed a research question addressing the review objectives. We further defined variables that could answer such questions, built a strategy to identify the related literature and proposed the instrumentation for data collection. All the three authors (i.e., A1, A2, and A3) actively participated in these steps.

Generalizability

Generalizability is concerned with the application of the results to different contexts and settings. Some of the identified methodological papers are also employed in other scientific domains e.g., business, and human sciences. However, they are not likely to be a comprehensive set of those domains, and therefore our results are not readily applicable to such contexts.

Furthermore, our results address multiple research methods, process phases, and maturity levels. However, the methodological support is not evenly distributed, and some gaps exist in relation to each category. We assume that this lack of equity reflects the actual state of the field. Thus, we suggest that further studies are needed to investigate the gaps and to provide support for all the research methods.

3.4 Results

Our systematic mapping identified 341 papers that present instruments or provide guidance to ESE research. An extensive list of the included references is
3.4 Results

provided in Appendix 3.8.4 and is available by means of the Web tool CERSE\(^2\). This Web tool also allows for researchers to filter the catalog references according to the paper categories, as described in Section 3.3.3.

With regard to their aims, most of the instruments (236 out of 341 included studies) describe guidelines, frameworks or experiences on how to conduct research. Assessment instruments, such as checklists and evaluation processes, are presented in 74 papers. Finally, circa of 42\% (41 papers) provide structures to classify and categorize the research. A number guidelines provide additional support to research-related activities, such as teaching and replication.

The included papers cover a broad spectrum of research methods (e.g., case study, experiment, SLR) and research phases (e.g., planning, execution, and analysis). The instruments are also classified according to their maturity. The categorization provides a structured map of the identified guidelines and supporting instruments, and also highlights existing gaps in the field.

Further, we detail our findings according to three distinct perspectives of categorization:

1. Research methods supported by the instrument;
2. Phases of the research process covered; and
3. Maturity with regarding the evaluation of the instrument.

3.4.1 Research Methods

In this section, we summarize the coverage of the research methods identified in our study. References to all identified studies according to the classification (i.e., research methods and aims), along with some highlights, are available in Appendix 3.8.2. A subset of these references can also be generated when selecting guidelines and other supporting literature using the CERSE Web tool\(^2\). The heuristics for guideline selection are detailed in Section 3.5.1.

**Time-trend.** Since 1980, several guidelines have been published, first addressing experimental studies, later on case studies, action research and other methodologies. An overview of the temporal trends of the identified literature concerning the different research methods is presented in Figure 3.2.

The time-trend plot shows when and how often a methodological support for each particular method in SE has been published. Experiment has been the research method with the highest number of methodological support, with 155

\(^{2}\text{http://www.cerse.org}\)
Figure 3.2: Time-trend of identified studies according to the topmost 10 cited research methods. The lines show the cumulative sum of methodological papers over the years i.e., each new publication pushes the related line up.

references (i.e., 43.4% of the total papers collected) since the first publication in 1980 [181]. Case study research comes second, with 76 references identified, and with the first guideline published in 1987 [142]. In relation to SLRs, we identified 70 papers published since 1997.
3.4 Results

One expects that the time-trend results are aligned to the needs for a specific SE-related methodology; however, the frequency in which such “primary studies” are conducted are not aimed by this study. It is as well likely that (a) primary research is conducted and reported after the methodological support in the context of SE is published; or (b) guidelines and supporting instruments emerges after primary studies have been published.

**Multiple classification.** A large part of the instruments target at more than one research method, often employing additional methods for data collection or data analysis. An example of such combinations is illustrated by the Venn diagram presented in Figure 3.3.

![Venn diagram](image)

Figure 3.3: Venn diagram shows the number of papers covering a set of research methods related to case study research. The diagram is divided into 15 separate sets, each of which presents the count of papers related to a particular research method or the intersection between them i.e., targeting more than one research method.

The Venn diagram shows the relations between case study research (CSR) and some other research methods often employed in such research. CSR often uses different research units to achieve data triangulation, an approach aimed at obtaining stronger evidence. From 72 guideline papers addressing CSR, more than half also discuss additional methods e.g., experiment, survey, and observation.
CERSE - Catalog for empirical research in software engineering: a systematic mapping study

Case Study Research

Case study research employs multiple methods of data collection to investigate a phenomenon in its natural settings [142]. The methodology is well suited for many SE research topics, as it addresses a contemporary case in depth. It aims to understand the particular case and create the basis for further research on the topic [391].

It is often associated with an element of triangulation i.e., employing different strategies (e.g., data collection instruments or sources) on the same target. Combining the results from a set of strategies can provide stronger evidence to support the findings that a single point of view.

Its flexibility implies a lack of boundaries as compared to other methodologies, and thus it is sometimes hard to classify case study research [391]. Both qualitative and quantitative approaches have been documented in the literature, and sometimes a mixed approach combines them both.

**Summary of findings.** We identified 76 papers supporting case study research in SE. Most of them presenting guidelines to conduct the process (58), but also instruments to classify (13) or evaluate the research (11).

Action Research

Action research aims to study a problem in the field by proposing, applying and evaluating a solution to solve it. The methodology is often described by the following phases: 1) diagnosing, 2) action planning, 3) action taking, 4) evaluating, and 5) specifying learning.

The researcher is an active part of the methodology, as he/she investigate the problem and propose action in response to it. Because of that, special measures to ensure the objectivity should be taken. Finally, knowledge is aggregated by collecting insights about the relationship between the intervention and the investigated phenomenon [449].

Action research is usually classified according to its focus and objective [112], namely, such as i) Action research focusing on change and reflection; ii) Action science trying to resolve conflicts between espoused and applied theories; iii) Participatory action research emphasizing participant collaboration; and iv) Action learning for programmed instruction and experiential learning.

**Summary of findings.** A total of 24 papers were identified as supporting action research in SE and related fields.
Design Science Research

While design science (DS) methodology shares the problem-solving paradigm of action research, it focuses on building and evaluating original artifacts (e.g., a model or a framework). These products are designed to meet the requirements of a particular problem and can be added to the shared knowledge base of researchers [440].

Summary of Findings Only 2 guidelines to conduct DS research were identified in our review. The first one [440] focuses on methodological guidelines considering nested knowledge and practice, two intrinsical facets of DS methodology. The other [240] presents a conceptual framework and seven guidelines for design science in IS research.

Interviews and Focus Groups

Interviews are often used in ESE to investigate the participants’ perception regarding the research topic. It usually involves a meeting between the researcher (a.k.a. interviewer) and the participant (a.k.a. interviewee) wherein the interviewer ask questions to collect in-depth qualitative data [243, 449]. Thus, interviews are a qualitative method, also defined as one of the field study methods i.e., featuring data collection outside a laboratory setting.

Interviews are classified according to the questions used [449], as such: i) Structured: mainly close-ended questions i.e., limiting the interviewee to a list of possible answer choices; ii) Unstructured: typically open-ended questions beginning with “what” and “how” to capture interviewee’s experience; and iii) Semi-Structured: combine both close- and open-ended questions.

Focus groups (also known as focused interviews or focus group interviews) allow the researcher to interact with a group of participants in a manner similar to the interviews [311]. Planning and execution phases are similar to unstructured interviews, but extra care should be taken to compensate its weakness e.g., secrecy and limited comprehension.

Summary of findings. We identified 15 papers presenting instruments to support interviews in SE. However, just two of those addressed interviews alone [243, 344]. Alternatively, the focus groups method is sometimes considered to be related to unstructured interviews [311].

Observation, Ethnography and Think-aloud

Participant observation is an empirical method to collect qualitative data in a less intrusive way than an interview. It consists in systematically inspect the
participant to capture behaviors and interactions that might not be noticed otherwise [398,449]. Along with the interview, observation is also considered as a field study.

According to the level of awareness of the observer, the research can be classified as [449]: i) Overt: the participant is aware of the process; and ii) Covert: the participant is not aware of the observation.

Ethnography is a methodology similar to the participant observation that requires the researchers to immerse themselves in the investigated environment [387]. The think-aloud protocol involves participants saying whatever comes into their mind (i.e., thinking aloud) as they are observed by researchers.

**Summary of findings.** Our mapping study identified 28 papers reporting guides and advice to the observation process. Similarly to the interviews, a significant portion of them relates observations to other research methods, such as case studies and action studies.

**Archival Research**

Archival research methods investigate the SE field by looking at their output and by-products e.g., source code, documentation, and reports [318]. It comprises activities such as locating the data, systematically collecting the data, analyzing and interpreting the data [449].

It is non-intrusive and requires little commitment from the subjects. However, real data may sometimes have missing components and/or be difficult to interpret [318]. Moreover, archival research can be a valuable data source for triangulation in both case study and action research.

Depending on the source of data, archival research can be classified as qualitative or quantitative. Qualitative data sources include e.g., meeting minutes, and software documentation, while quantitative data come from e.g., project databases, software repositories, change logs, and fault reports.

Archival research is often performed in industrial settings, but may also be applied over academic papers. According to the decision-making structure presented in [449], systematic literature reviews are archival research applied to academic literature. However, we treated SLRs separately (Section 3.4.1), as the guidelines for secondary studies contain specific steps related to study identification, selection, etc.

Postmorten reviews investigate existing data from concluded projects to gather insights for future cases. These light studies employ retrospective reflection often allied with support from individuals that participated in the project.
3.4 Results

Summary of findings. 13 articles related to archival research were collected in our review. None of the papers is focused on archival research alone, and the advice is often found in field study methodology.

Systematic Literature Review

Systematic Literature Review (SLR) aims to collect and summarize the available literature regarding a particular investigated topic of the phenomenon. It employs a rigorous and systematic methodology to retrieve a wide range of pieces of research, aggregating their results to draw more general conclusions than individual studies in isolation [294]. The systematic process is often described in three main phases, and its particular activities or stages:

- **Planning**: specify a research question and develop a research protocol, which details the process used to conduct and report the review;
- **Conducting**: conduct a search strategy to retrieve candidate studies that are further selected (to remove irrelevant studies) and assessed (to ensure the quality of evidence); later, data relevant to answer the research question are extracted, aggregated and synthesized;
- **Reporting**: write and validate a report given the target audience and appropriated format.

SLRs are often described as secondary studies, a general category that also includes systematic mapping studies (see Section 3.4.1). SLRs provide a more in-depth analysis of the selected studies than mapping studies, but often require more effort [361].

Summary of findings. A total of 70 papers provide guidance and support to SLR, 30 of them are also related to systematic mapping studies.

Systematic Mapping Study

Systematic mappings are secondary studies that aggregate the available literature in a more comprehensive and often visual summary. The differences from other secondary studies (i.e., SLRs) also reflect on the process employed, as mapping studies often require less effort. The quality assessment activity is not required, and both study selection and data extraction only consider high-level characteristics of the identified papers [361].

Summary of findings. We identified 35 papers focused systematic mapping studies, most of them also addressing SRLs. Only 5 papers provide specific guidance to systematic mapping alone.
Survey Research and Delphi Method

Surveys aim to collect and summarize evidence from a large representative sample of a population of interest. Outcomes from survey research are then combined to identify patterns with the aim of generalizing to the overall population. Identified patterns can be even compared to different populations and evaluated over time.

According to the research purpose, survey research can be: i) exploratory: whether aiming to identify preliminary concepts, or discover new dimensions from the population of interest; ii) descriptive: whether investigating the opinions or phenomenon occurring in the population or a subset of the population; or iii) explanatory: whether testing theories and causal relations.

Regarding the data collected, survey research can also employ a qualitative or quantitative process, or even a mixed research process when collecting simultaneously qualitative and quantitative data.

Delphi method is a survey-based communication technique to obtain the consensus of a group of experts. Through several rounds, the group discusses a problem and answers questionnaires related to it. An anonymous summary of the responses is provided at the end of each round, thus promoting further discussion and consensus.

Summary of findings. In total, 39 papers supporting the survey process and its activities were identified. Some of them are focused on specific aspects of survey research, as others also cover it along with other field study methods, e.g., observations, interviews, and case study research.

Experiments and Quasi-Experiments

Experimental research is a vast though well-explored field on ESE often used to compare two or more alternative treatments (e.g., methods, techniques or tools). Participants are assigned and subject to a treatment condition, and the effects of such are recorded. Further, the data between the treatments are analyzed and compared through statistical methods.

Experimental design is often classified according to its characteristics, such as: i) laboratory or field setting: whether conducted in a controlled or real-world environment; ii) controlled or non-controlled: whether or not employ controls i.e., individuals not subject to any treatment; iii) randomized or non-randomized: whether randomly assign the participants to the treatment or not. iv) laboratory or quasi-experiments: randomized controlled trials (RCT) are con-
ducted in a laboratory setting, whereas quasi-experiment designs (a.k.a. natural experiments) are carried out in a field setting.

**Summary of findings.** Most of the literature identified in our review is related to experiments i.e., 155 out of 341 papers. Very often identified papers relate experiments to case study research, and to observation, survey and simulation methods.

**Simulation**

Simulation-based studies (SBS) use the model of a real-world entity (e.g., a process or product) to evaluate alternative solutions under investigation. The model can reproduce the components of an actual entity in a virtual state, with varying degrees of accuracy [449]. SBS share several characteristics with the controlled methods for technology evaluation, such as experimental designs [456].

According to the model, simulations can be characterized as: i) *static or dynamic*: the system is time dependent or not; ii) *discrete or continuous*: the investigated events are countable or continuous in time, and iii) *stochastic or deterministic*: the simulation depends on probabilistic behavior or not. Moreover, the employed research logic also depends on the model: i) *Deductive*: generated from theory and compared to actual observations, and ii) *Inductive*: derivative from gathered data and used to examine the relationships between its components.

**Summary of findings.** Our review identified only 5 papers related to simulation, most of which are also relevant for experimental research.

**Thematic Analysis**

Also known as thematic synthesis, it is another qualitative data analysis technique used in SE to provide understanding about an investigated topic. It identifies the recurring concepts from multiple data sources exploring the relationships between them to create a model of high-order themes [178].

Although some differences in the literature [178, 449], the process can be described by the following phases: i) initial reading/familiarizing of data/text; ii) identifying initial codes/segments of text; iii) labeling/identifying themes; iv) reviewing themes/reduce overlaps; v) defining the results.

Although very similar to thematic analysis, content analysis is focused on counting the frequency of concepts within data. It requires well-defined categories and strict rules, and often employ computational tools to aid the process.
Thematic analysis can be classified according to the process to identify the themes as i) semantic: when based on explicit meaning in the data; and ii) latent: based on underlying ideas within the data that is theorized. The semantic approach tends to be more positivist whereas the latent approach is often interpretivist.

Summary of findings. From 5 papers identified, only one is focused on supporting thematic analysis. Other papers are concern overview various data analysis techniques (e.g., [179,231]) or how to combine various data collection and data analysis methods (e.g., [210,327]).

Grounded Theory

The Grounded Theory Method (GTM) is intended to construct a theory from a systematic process of data analysis. It has been used to understand how participants deal with problems concerned to them, based on data gathered from opinions and behaviors [121].

GTM is an open-ended inquiry process that involves collecting, coding and categorizing qualitative data (e.g., interview transcripts, literature review, or archival data) until a saturation level is reached. Potential relationships between the codes are further interpreted to build theories and interrelated hypotheses. The theories can then be tested by a further study [449].

Two different versions of GTM exist, according to the data coding and interpretation: i) Glaserian: identifying participants’ perspectives data at an abstract level, and later conceptualizes them to find patterns, and ii) Straussian: develop categories according to the literature or the researcher’s experience simultaneously to the coding process.

Summary of findings. 8 identified papers presented support to grounded theory process. Half of them are focused on grounded theory solely, while as four other include other field study or data analysis methods.

Hermeneutics

A qualitative data analysis approach focused on the interpretation of meanings. The method requires interpreting small parts of the data that leads to an understanding of the whole. It is especially suitable for human behavior studies, including information systems and software engineering research [449].

Hermeneutics most fundamental principle is called circle of understanding (a.k.a. hermeneutics circle). The process uses a dialogic approach to interpreting new knowledge connected to prior beliefs. In that sense, a global understand-
3.4 Results

ing of the context can lead to an improved understanding of the parts, and vice versa [309]. Hermeneutics is often characterized by philosophical perspectives applied i.e., conservatism, pragmatism, criticism, the radicalism [163].

Summary of findings. We identified just two studies presenting an overview of hermeneutics. They detail the concepts and principles of the approach, discussing its philosophical aspects. Both papers (i.e., [163, 309]) outline the application through interpretive studies on the information systems field.

Statistical Analysis and Meta-Analysis

Statistical methodologies have been used for data analysis in a multitude of fields. The statistical approach relies on the application of mathematical techniques (e.g., linear algebra, differential equations, and probability theory) to interpret quantitative data regarding a population or a model to be studied.

Although quantitative data (e.g., collected in an experiment or survey-based research) is required at statistical analysis; qualitative data, such as concepts and themes, can be converted to quantitative form by e.g., counting the frequency of appearance.

On the one hand, descriptive statistics is meant to summarize data by describing, aggregating and presenting association between the variables. On the other hand, inferential statistics includes hypotheses testing, regression analysis, and estimation through data mining techniques [449]. Moreover, according to the distribution of data, statistical methods can be i) parametric: when making assumptions on the distribution of data (e.g., normal distributed); and ii) non-parametric: no requirements on probabilistic distribution, but still bounded by scale, amount of groups investigated, etc.

Meta-analysis is a statistical procedure that combines and summarizes the results of multiple individual studies.

Summary of findings. Our review identified 17 papers related to the application of statistical analysis in SE studies. Most of them (11) are related to experimental data collection, and just a few of them (2) also address survey-based research.

3.4.2 Research Process’ Phases Covered

Table 3.3 shows the number of identified studies associated with a research method, as discussed in Section 3.4.1. Identified studies do not equally cover all the research phases, such as: i) Planning; ii) Execution; iii) Analysis; and iv) Reporting. These phases represent the conventional life cycle of typical
empirical research; however, more specific discrete activities could be required according to the method used (e.g., to write a review protocol during the SLR planning).

Table 3.3: Number and percentage of studies according to research methods and covered phases.

<table>
<thead>
<tr>
<th>Method</th>
<th>Planning</th>
<th>Execution</th>
<th>Analysis</th>
<th>Reporting</th>
<th>All Phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Study</td>
<td>60</td>
<td>47</td>
<td>40</td>
<td>33</td>
<td>18</td>
</tr>
<tr>
<td>Action Research</td>
<td>17</td>
<td>16</td>
<td>10</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Design Science</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Interviews</td>
<td>10</td>
<td>13</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Focus Group</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Observation</td>
<td>14</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Ethnography</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Think-aloud</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Archival Research</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Postmortem Review</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Systematic Literature Review</td>
<td>34</td>
<td>54</td>
<td>33</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>Systematic Mapping</td>
<td>14</td>
<td>29</td>
<td>13</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Survey</td>
<td>32</td>
<td>28</td>
<td>18</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Delphi Method</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Experiment</td>
<td>89</td>
<td>85</td>
<td>80</td>
<td>59</td>
<td>19</td>
</tr>
<tr>
<td>Quasi-Experiment</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Simulation</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Thematic Analysis</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Content Analysis</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grounded Theory</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hermeneutics</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>7</td>
<td>5</td>
<td>13</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Meta-Analysis</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Overall empirical research</td>
<td>27</td>
<td>17</td>
<td>9</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>54.9%</strong></td>
<td><strong>60.2%</strong></td>
<td><strong>47.6%</strong></td>
<td><strong>31.9%</strong></td>
<td><strong>14%</strong></td>
</tr>
</tbody>
</table>
Most of the papers address multiple phases, sometimes providing support for the entire research process. The number of papers that cover all the four phases is presented in the last column. 14% of the identified studies cover all the phases of the related process. Most of the papers address the execution (60.2%), planning (54.9%) and analysis (47.6%). Even few papers (31.9%) provide guidance to report the research.

Although those methodological papers are often not covering all the research phases, they can be particularly useful for researchers willing to conduct more complex studies i.e., combining research methods. On the one hand, some of those papers explain how different methods can be combined, using the results from previous studies to more in-depth research. On the other hand, others discuss the definitions of how research methods are distinct, helping researchers in selecting a more suitable setting for an intended research goal. A comprehensive list of studies addressing more than one research method is given in Appendix 3.8.3.

3.4.3 Validated and Evaluated Guidelines

Some of the guidelines, assessment instruments, and knowledge organization systems have been validated (67 papers) or evaluated within its actual context (37 papers). Herein validation and evaluation research (see a glossary of terms in Appendix 3.8.1) mean distinct degrees of maturity of an instrument according to empirical investigations employed to assess them.

Validation and evaluation are not equally distributed among all the research methods, as shown in Figure 3.4. SLR and systematic mapping studies are the ones presenting more evaluation and validation (i.e., considering the total amount of papers), but also experiment and case study research have been evaluated in the SE domain. Several research methods are targeted of validation studies, but no evaluations, such as action research, grounded theory and content analysis. Finally, some methods (e.g., think-aloud, content analysis and hermeneutics) presented a relatively higher number of validated or evaluated papers, but this is mainly a result of few identified studies. Some of the highlighted references are given below:

Overall empirical research. Most of the literature aiming at the overall empirical research (i.e., rather than a particular research method) is in the early stages of maturity (i.e., opinion and philosophical papers). Due to the high level of abstraction of such philosophical papers, it is hard to evaluate their support for research and practice. Three distinct KOS [148, 241, 257] were validated through their application in a series of reported studies.
Figure 3.4: 100% stacked column chart of identified papers according to maturity level. Each column is split into colored bar segments representing the relative contributions (i.e., in percentage) to the total bar. The total number of papers identified for each column is shown between parenthesis (e.g., there are a total of 76 case study related papers). Validated and evaluated papers are highlighted in orange and green, respectively.

**Case Study Research.** three instruments supporting CSR were evaluated: i) DESMET has been subject to some evaluation exercises [285], even using the methodology as the evaluation instrument itself [276]; ii) a longitudinal case study methodology is both presented and evaluated by Laurie McLeod and MacDonell [327], and iii) a unified checklist for overall empirical research [446] has been evaluated by Condori-Fernandez et al. [173] for conducting case studies
3.4 Results

and experiments. Additional papers (e.g., [218, 391]) also presented validation studies.

**Action Research.** We found no evidence of any evaluations. However, three papers addressing AR were validated, two of which are KOS [396, 452] and another one presenting a comprehensive set of guidelines and an assessment checklist for case study and action research [391].

**Interviews and Observations.** No paper addressing those methods alone presented evaluations. However, guidelines addressing more general research methods, such as case study [327, 391] and experiment [201], while employing interviews and observations were proposed and evaluated through empirical.

**Secondary Studies.** Kitchenham and Charters’ SLR guidelines [277, 294] were evaluated by comparison with other domains [155], by interviewing practitioners [128] and by a systematic review process itself [134]. The EBSE approach [295] was also validated [258, 349, 376, 377]) and evaluated [378] by applying it with students.

Several activities of secondary studies (i.e., both SLRs and mapping studies) has been evaluated, consist of a plethora of best practices reported. The most investigated topic is the search strategy (e.g., [202, 249, 282, 291, 447, 461]), covering the differences between manual search (and snowballing techniques) and database search (sometimes through automated mechanisms). Other topics include research question formulation [183], study selection [122, 158, 159, 429], quality assessment [290], data extraction [429], data analysis [198], software tool support [324, 325], and repeatability [281].

**Survey Research.** No evaluation of guidelines and assessment instruments for survey research were identified. Just two papers addressing survey research were validated, both also covering case study and experiment research [218, 452].

**Experiments.** According to Dieste *et al.* [199], the quality of experiments is related to the degree of use of methods that reduce bias. Several assessment instruments addressing different criteria for experiment were both proposed and evaluated [173, 261, 270, 276, 278, 279, 285, 289]. Another metric for experiment’s assessment is the replicability; which relates to two specifically evaluated guidelines [392, 425].

**Simulation.** Identified literature regarding simulation studies showed no evaluation. However, an assessment instrument for simulation and experiments [456] was validated. The instrument aimed to improve simulation-based studies understandability, replicability, generalization, and validity.

**Thematic Analysis and Grounded Theory.** We found a single evaluated paper addressing the thematic synthesis in SE research [327]. None of the identified papers presented an evaluation of the GTM process in the SE con-
text. However, two studies addressing different data analysis methods, such as thematic analysis and grounded theory (i.e., [179, 231]) discuss the difficulties in applying such technique in SE studies.

**Statistical Analysis.** Kirsopp and Shepperd [275] evaluated the lack of accuracy of prediction statistics for effort estimation with very small samples. Moreover, both a guideline paper [327] and an assessment instrument [261] addressing experimental studies employing statistical analysis were also validated.

### 3.5 Discussion

The systematic mapping detailed herein provided evidence used to build a catalog of guidelines for empirical research in software engineering. This extensive set of references is structured according to the target research method and type of instrument described. To summarize the included studies, we also collected additional information related to the identified instruments, such as which research methods are covered and which phases are targeted, its main objectives (e.g., conduct, evaluate or classify the research) and the maturity level of the instrument.

#### 3.5.1 Heuristics for Guideline Selection

The results of this mapping study are aimed to support researchers in identifying and selecting guidelines and other supporting literature (i.e., assessment instruments and knowledge organization systems) suitable for their research. The catalog presented herein is the first step towards achieving such aim. It provides a comprehensive set of methodological studies, classified according to their related research methods.

Researchers willing to use it effectively still need to retrieve a subset of the catalog and to form an opinion about which of guidelines are most suitable for their research. Therefore, we proposed a practical approach on how to find a subset based on the characteristics of the proposed research, as illustrated in Figure 3.5, and described below:

1. **Define a research proposal.** A research proposal should be founded on underlying theories, implemented through one or more related studies, aiming application or derivation of new theories. Before proceeding to the next phase, we suggest outlining the individual studies that compose the research proposal. Researchers willing to read about writing good research proposals can benefit from reference books e.g., [103, 116].
<table>
<thead>
<tr>
<th>Steps</th>
<th>Actions</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Define a research proposal</td>
<td>Identify individual studies that compose the research proposal</td>
<td>Creswell [103], Shull et al. [116], etc.</td>
</tr>
<tr>
<td>2. Identify the research units</td>
<td>Select the research method(s) for each individual study</td>
<td>Wohlin and Aurum [449] framework</td>
</tr>
<tr>
<td>3. Choose the suitable references</td>
<td>Filter methodological supporting instruments by:</td>
<td>CERSE catalog of empirical research for software engineering and web tool (<a href="http://www.jeffersonmoller.com/cerse">http://www.jeffersonmoller.com/cerse</a>)</td>
</tr>
<tr>
<td></td>
<td>- research method(s)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- process phase(s)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- type/objective of instrument</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- maturity level</td>
<td></td>
</tr>
<tr>
<td>4. Aggregate the supporting literature</td>
<td>Combine a reference list for further consult</td>
<td></td>
</tr>
<tr>
<td>5. Reflect on the experience</td>
<td>Follow the guidelines and report lessons learnt</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.5: Heuristics for guideline selection. Each step requires specific actions and can be linked to particular literature. The initial steps (i.e., 1 and 2) are addressed by related work, whereas the intermediate steps (3 and 4) are supported by the catalog of empirical research methods.
2. **Identify the research units.** Those units include the research methods intended for each individual study. Note that the choice of a particular methodology affects the data collection and data analysis techniques. The decision point based structure proposed by Wohlin and Aurum [449] could support this particular step.

3. **Select the suitable references.** Search in the catalog (Appendix 3.8.2) or the CERSE Web tool (http://www.cerse.org) for the section related to each of the research methods intended for use. Often, an extensive list of available guidelines is presented. It should be necessary to filter the available guidelines according to specific needs. Note the following heuristic approach to better choose the guidelines:

   3.a. **Guidelines for conducting the research:** Prefer the guideline papers that address the whole process. If there is no such guideline, build a set of papers that covers all the intended phases.

   3.b. **Context of the research:** As some research methods present several available guidelines (e.g., case study, survey, experiment), we suggest researchers to screen the list keeping in mind the particular context of the proposed research. Moreover, the discussion in section 3.4.1 include some highlighted references that could be valuable.

   3.c. **Multiple research methods:** In case that your project employ a set related studies, also browse the guidelines addressing those research methods together (Appendix 3.8.3).

   3.d. **Mature guidelines:** Give priority to evaluated (or otherwise validated) guidelines (as shown in Section 3.4.3).

   3.e. **Assessment instruments:** Add to your references the assessment instruments or checklists to assure that the planning and reporting meets the quality requirements. Note that some guidelines already include quality assessment steps.

   3.f. **Knowledge organization systems:** If necessary for planning or reporting, also read KOS to support the methodology categorization.

   3.g. **Informed decision:** Finally, it is essential to assess the resulting references in order to decide which guidelines to use. Consider the limitations, strengths, and challenges of each instrument with regard to the proposed research. The highlights provided in Appendix 3.8.2 are intended to support this step.
4. **Aggregate the supporting literature.** Combine the references into a joint list that could be easily consulted. Make sure that the guidelines are followed during the project, and any change is reported.

5. **Reflect on the experience.** Finally, it is desirable to conduct a critical analysis on the use of the selected guidelines and other supporting instruments. Such reflexion could be included in the study report, thus informing fellow researchers the lessons learned during the process.

Novice researchers are encouraged to follow the five-step strategy, making use of supporting references listed on the right hand side of Figure 3.5 to design a research proposal. More experienced researchers (i.e., already used to the initial steps) would most likely only benefit from steps 3 and 4. Furthermore, researchers still need to make an informed decision on the relevant guidelines to use, as the subset is likely to contain overlapping or contextual-related guidelines.

### 3.5.2 Usage Scenario for Guideline Selection

Based upon the heuristics abovementioned, we can derive a usage scenario rooted on an exemplar of ESE research proposal originally described in [449]. The research scenario is as follows:

*Example.* A Student wants to conduct research on technical debt, as this phenomenon has not been addressed enough in academic literature, even though practitioners have emphasized its importance. (...) The student believes that both academia and practitioners will benefit from having a framework that defines what constitutes technical debt, why it occurs and its detrimental implications in projects. Practitioners will be involved to evaluate the framework.

1. **Define a research proposal.** In the excerpt above, the Student describes the research idea, the intended contributions and the target group. In the first step of the proposed heuristics for guideline selection, the Student should also identify the individual studies needed to achieve the aims of the research. In this context it is important to explicitly highlight the methods to be used in the different studies. We already highlighted the important information needed to identify the research units.

*Example.* (...) Student conducts a systematic literature (M1), using thematic analysis (M2) to identify to what extent technical debt is covered in academic literature. The outcome of the systematic
literature is a framework that illustrates reasons and outcomes of technical debt. Then the student carries out a multi-vocal literature review (M3) to revise the framework. The multi-vocal literature review (M3) considers gray literature authored by practitioners, which is considered as a data-point of a case study (M4). Both the systematic literature review and the multi-vocal literature review apply a qualitative research approach, (...) and use thematic analysis as data analysis method (M2). The resulting framework is evaluated by interviewing several practitioners (M5). The interviews are transcribed and thematic analysis (M2) is applied. Then, experts evaluate the framework and reviews are integrated to research findings.

2. Identify the research units. The excerpt above already mentioned several methods in the research proposal. Overall, the student conducts three different research units: (RU1) a preliminary literature review, followed by (RU2) a case study, and further (RU3) an interview study. According to the proposal, the following methods should be looked up in CERSE:

- RU1: makes use of the methods systematic literature reviews (M1) and thematic analysis (M2).

- RU2: makes use of the methods “multi-vocal” literature review (M3), which uses the publications of practitioners as data for a qualitative data analysis (M2). The process for data collection described in the example can be considered as an archival research method (M3) following the definition by [449] which classifies the multi-vocal literature reviews as such. In summary, RU2 employs the methods case study (M4) research, archival research (M3) and thematic analysis (M2).

- RU3: uses interviews (M5) and also thematic analysis (M2).

After having identified the relevant methods for the research units, the student should consult the catalog (see Appendix 3.8.2 or the CERSE Web tool (http://www.cerse.org) and select the research methods that fit each individual unit.

3. Select the suitable references. Now, as a starting point, the student searches the catalog for five methods identified from the research proposal, and 152 references of supporting instruments are found. Table 3.4 maps the relevant parts of CERSE that should be consulted.

The Student should filter the results further in order to narrow down a set that is more appropriate to handle. The heuristic for the selection of the relevant
Table 3.4: Candidate references for the exemplar scenario.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference tables (Appendix 3.8.2)</th>
<th>Qty.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M1) SLR</td>
<td>Section 3.8.2.11</td>
<td>69</td>
</tr>
<tr>
<td>(M2) Thematic Analysis</td>
<td>Section 3.8.2.18</td>
<td>5</td>
</tr>
<tr>
<td>(M3) Archival Research</td>
<td>Section 3.8.2.9</td>
<td>12</td>
</tr>
<tr>
<td>(M4) Case Study</td>
<td>Section 3.8.2.1</td>
<td>76</td>
</tr>
<tr>
<td>(M5) Interviews</td>
<td>Section 3.8.2.4</td>
<td>14</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>152</strong></td>
</tr>
</tbody>
</table>

guidelines (see Section 3.5.1, step 3.a-3.g) are used to guide the selection. Table 3.5 details the selection criteria and the rationale for each step.

4. **Aggregate the supporting literature.** Later, the Student selects the included candidates, and aggregates them into a list (illustrated in Table 3.6) of supporting literature, as follows:

**RU1:** Considering the support to all the phases of the research process (*Heuristic 3.a*), the student could benefit from the guidelines for performing SLRs [294]. This paper is the basis for the student’s literature review. Moreover, by checking the references employing both the SLR and thematic synthesis, the student could identify two additional studies [210, 231] discussing a combination of those methods (*Heuristic 3.c*).

**RU2:** Case study research is covered by a long list of guidelines, thus making harder for the Student to decide. The guidelines for conducting and reporting case study research in SE [391] is a well established reference (*Heuristic 3.g.*). The paper presents recommendations for all the phases of the research process (*Heuristic 3.a.*) and was also validated (*Heuristic 3.d.*). Further, archival research as part of case study research is also supported by Runeson and Höst [391] (*Heuristic 3.c.*).

Alternatively, the DESMET approach [300] is mainly developed to promote a case study approach to evaluate SE methods and tools. However, as the evaluations proposed by DESMET are mainly quantitative, they do not match the context of the research, therefore excluded from the candidates list (*Heuristic 3.b.*).

**RU3:** Interviews with practitioners are planned to evaluate the proposed framework. Advice on how to conduct semi-structured interviews are given by Hove and Anda [243] (*Heuristic 3.a.*). Moreover, thematic analysis is also used to analyze the data collected during interviews. In that sense, additional support
Table 3.5: Selection criteria for the exemplar scenario

<table>
<thead>
<tr>
<th>Step</th>
<th>Criteria</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.a.</td>
<td>Type of Instrument = <em>Guidelines</em></td>
<td>The student’s main aim is to identify instruments that help conduct the research according to the five identified methods.</td>
</tr>
<tr>
<td>3.a.</td>
<td>Process phases = <em>All</em></td>
<td>It is also important for the Student to identify guidelines that address all the phases of the process.</td>
</tr>
<tr>
<td>3.b.</td>
<td>Context of the research</td>
<td>Instruments that provide guidance to how to carry out practitioner-oriented research in the context of technical debt are preferred. Moreover, the research is mainly exploratory and qualitative by nature.</td>
</tr>
<tr>
<td>3.c.</td>
<td>Multiple research methods</td>
<td>Aiming to find guidelines that provide more complete methodological support to conduct the entire research proposal, the Student selects mainly guidelines that combine support to more than one method for the same unit of research.</td>
</tr>
<tr>
<td>3.d.</td>
<td>Maturity = <em>evaluated, validated</em></td>
<td>If two or more guidelines addressing the same method are available, the Student should select one that has been evaluated or validated.</td>
</tr>
<tr>
<td>3.e/3.f.</td>
<td>Type of Instrument != <em>Ass. instruments, KOSs</em></td>
<td>At this stage of the research, the Student is not interested in supporting instruments other than guidelines. Assessment instruments can be valuable in the future to assess if the research was conducted according to quality standards.</td>
</tr>
<tr>
<td>3.g.</td>
<td>Decision-making</td>
<td>Finally, the Student should critically read and assess the candidates to make an informed decision as to which guidelines to select. The Student also makes use of the highlights in Appendix 3.8.2 to support this process.</td>
</tr>
</tbody>
</table>
### 3.6 Conclusions

In this Chapter, we detailed existing guidelines, assessment instruments and knowledge organization systems to empirical research in SE. Via a systematic mapping 341 references from papers proposing such methodological support were gathered and selected. Some relevant findings were described in detail in section 3.4.

---

#### Table 3.6: List of supporting literature for the exemplar scenario

<table>
<thead>
<tr>
<th>Res. Unit</th>
<th>Reference</th>
<th>Selection criteria</th>
<th>Research Method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU1</td>
<td>Guidelines for performing Systematic Literature Reviews in Software Engineering [294]</td>
<td>3.a, 3.g</td>
<td>M1</td>
</tr>
<tr>
<td></td>
<td>Applying Systematic Reviews to Diverse Study Types: An Experience Report [210]</td>
<td>3.c</td>
<td>M1+M2</td>
</tr>
<tr>
<td>RU2</td>
<td>Guidelines for conducting and reporting case study research in software engineering [391]</td>
<td>3.a-3.d, 3.g</td>
<td>M3+M4+M5</td>
</tr>
<tr>
<td>RU3</td>
<td>Experiences from Conducting Semi-Structured Interviews in Empirical Software Engineering Research [243]</td>
<td>3.a, 3.g</td>
<td>M5</td>
</tr>
<tr>
<td>RU1-3</td>
<td>Recommended Steps for Thematic Synthesis in Software Engineering[178]</td>
<td>3.a, 3.g</td>
<td>M2</td>
</tr>
</tbody>
</table>

---

For interviews is provided by guidelines already in the student’s list (Runeson and Höst [391], *Heuristic 3.a*).

5. **Moving forward.** The Student should now become familiar with the selected guidelines and make sure that they are followed during the research process. Any changes to the process should be reported in order to allow its replication or validation. After conducting the research, the Student should also report the experiences with the chosen methods and selected guidelines, providing insights on their usage to the research community.
3.6.1 Contributions of the Mapping Study

The mapping study’s main contribution is to provide a list of guidelines, assessment instruments, and KOS supporting different research methods in ESE. The catalog is further provided in Appendix 3.8.2. Furthermore, we summarize the results according to three different perspectives:

- **Research method.** A comprehensive list of covered research methods and some highlighted references are further discussed in Section 3.4.1. Further information on papers covering more than one method is given in Appendix 3.8.2. We also note that some of the methods are inter-related to another one, sharing some similar characteristics.

- **Phases of the research process.** Most of the identified papers provide methodological support for more than one phase of the process. However, just a few of them (15%) covers all the phases together, as shown in Section 3.4.2. We suggest that researchers willing to employ such guidance carefully consider to aggregate a set of papers to bridge such gaps.

- **Evaluated instruments.** Evaluation studies are important to assess the completeness and accuracy of such methodological supporting literature in ESE research. Our findings show that more mature research (i.e., evaluated guidelines and assessment instruments) are 10% of the identified studies. Moreover, there is a contrast in relation to the extension in which the papers addressing different research methods have been evaluated (also validated) in SE domain.

3.6.2 Implications for Research

Software engineering researchers are the main recipients of this work. Foremost, the catalog provides a comprehensive collection of the research methods employed in ESE research. Although similar inventories have been provided in some other papers (e.g., [148, 449, 456, 457]), our actual list is mapped to the guidelines, assessment instruments and knowledge organization systems reported in the SE literature.

**Opportunities for Research.** We encourage further research aiming the gaps identified in the Section 3.6.1 that are worthy of further investigation:

- **Research methods are not well-supported by specific guidelines for the software engineering domain e.g., design science, focus groups, think-aloud protocol, Delphi method.** By reflecting on how the guidelines from different domains could address SE research problems and conducting studies to validate such approaches, the community can improve the available guidelines with experience reports and lessons learned.
3.6 Conclusions

- **Guidelines and supporting instruments often do not cover all the research phases together.** We should be able to address this gap by aggregating and comparing the relevant guidelines and instruments related to a specific research method or context.

- **Supporting instruments are scarcely evaluated.** Results from evaluations can increase the reliability on using that methodological guidance, and identify improvement needs. By improving the guidelines, the research community can also increase the quality of empirical studies conducted using such guidelines as basis [421]. We also noted the need to conduct evaluation studies, especially on the research methods still in the early stages of maturity.

**Learning Aspects.** Moreover, the catalog can be employed by students and novice researchers learning how to conduct empirical studies. By mapping the catalog references to the particular units of a research proposal, researchers can select a subset of the guidelines fitting their intended research processes. A strategy to this is described in Section 3.5.1 and exemplified through a usage scenario in Section 3.5.2. The heuristics alone are not intended to guide novices through the study design phase, but instead can be employed together with other pedagogical instruments (e.g., [115, 449]) for learning or teaching purposes.

3.6.3 Future Work

As future work, we plan to update the catalog periodically by applying the search strategy on the most recent journal editions and conferences proceedings, as well as also conducting further snowballing iterations. The updates should be made available through the Web tool’s database. We also aim to conduct an empirical investigation to validate the proposed catalog and the related heuristics for guidelines selection.

In addition to that, we intent to conduct further studies aiming the relation between the methodological support and the empirical research in SE. In particular, we aim to investigate whether the availability of a specific SE-related methodology is aligned to the quantity and quality of primary studies using such method.
3.7 References


*(See additional references from the mapping study in Appendix 3.8.4)*
3.8 Appendix

3.8.1 Glossary

Maturity levels / score

Evaluation study An empirical research that investigates the implementation of a guideline or supporting instrument in real life setting e.g., with actual researchers and/or using the research projects conducted by them. Evaluated research is the highest stage of maturity according to the classification of research facets [445].

Validation study An empirical study aiming to validate a guideline or supporting instrument in a laboratory setting, or otherwise using toy examples. Validation research is weaker than evaluation studies, according to the classification of research facets [445].

Classification of supporting instruments

Assessment instrument Methodological tools to ensure that the research process is followed according to particular evaluation criteria. Examples of such tools include checklists, rubrics and assessment metrics [209,421]. Some assessment instruments rather aim at the reported study or the guidelines supporting the research.

Guidelines A step-by-step structure that details the process of conducting research according to a particular methodology [119,421]. It is often described as regarding the process phases and their underlying actions. In this chapter, we also relate the guidelines to frameworks and tutorials that provide guidance to how the research should be executed. Some guidelines also address how to teach research, and how to replicate the research.

Knowledge organization system A semantic structure that formally describes the research process via vocabulary control and knowledge organization [118]. Examples of KOs are taxonomies, ontologies and classification schemes.
### 3.8 Appendix

#### 3.8.2 Catalog of Empirical Research

##### 3.8.2.1 Case study

<table>
<thead>
<tr>
<th>Phase</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[139, 152, 171, 302, 304, 323, 364, 443, 455]</td>
<td></td>
<td>[171, 206, 219, 364, 443]</td>
</tr>
<tr>
<td>Execution</td>
<td>[151, 373, 407]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>[180, 215, 350, 372, 430]</td>
<td>[215]</td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>[363, 366, 444]</td>
<td>[276]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[254, 288, 303, 316, 318, 337, 374, 382, 395, 435]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td>[400]</td>
<td>[456]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[130, 163, 164, 220, 296, 301, 306, 309, 412, 413, 420, 446]</td>
<td>[446]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>[165]</td>
<td>[452]</td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td>[362]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis + Reporting</td>
<td>[307]</td>
<td>[399]</td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[124, 142, 169, 184, 217, 218, 242, 252, 299, 308, 327, 353, 359, 384, 391, 416, 421]</td>
<td>[142, 242, 285, 299, 391]</td>
<td>[142, 218]</td>
</tr>
</tbody>
</table>

#### Highlights

1. **Guidelines for case study research**: Several guidelines are intended to guide researchers through all phases of case study research. They often include recommended practices to report and review such studies. Those instruments include guidelines and tutorials for conducting CSR in SE (e.g., [242, 359, 391]) and IS [142].

   - **Quantitative case studies**: The DESMET series [107] propose a methodology for evaluating software engineering methods and tools through quantitative data collection and feature analysis. Particular advice includes selecting host projects, finding proper metrics, minimizing confounding factors, monitoring data collection, and quantitative analysis.
CERSE - Catalog for empirical research in software engineering: a systematic mapping study

- *Mixed case studies*: A unified checklist [173,446] proposes a research design based on the relationship between case studies and experimental research methods.

- *Triangulation*: The use of multiple data sources as a means to identify more and more trustful findings is discussed by Bratthall and Jørgensen [151].

- *Replication*: Some lessons learned on the replication of case study research are given in [337,382]. Several levels of replication are categorized and discussed by Ohlsson and Runeson [350].

2. *Assessment instruments*: Several criteria can be used to evaluate CSR, such as design and report phases [173,391], validity threats [215,362] and generalization of claims [399].

3. *KOS*: The main characteristics of case studies and experiments are presented by [364]. Some additional schemes propose the categorization based on the differences between positivist and interpretative approaches [206].

### 3.8.2.2 Action Research

<table>
<thead>
<tr>
<th></th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[139,171]</td>
<td></td>
<td>[171]</td>
</tr>
<tr>
<td>Reporting</td>
<td></td>
<td>[396]</td>
<td>[185]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[127,204,315]</td>
<td></td>
<td>[318]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[141,326,341]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>[314]</td>
<td></td>
<td>[452]</td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td>[362]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[140,142,217,391,416,421]</td>
<td>[142,391]</td>
<td>[142]</td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for action research in IS*: Our review identified no instrument to conduct action research according to particular software engineering needs. Most of the guidelines cited in SE domain come from information systems (e.g., [140,185,217,315,341]).

   - *Action and case studies*: Action research is sometimes defined as action case studies, as it shares several characteristics with case studies.
3.8 Appendix

Some papers present guidelines along with case study research (e.g., [142, 171]). One particular article discusses the contrasts between them [139].

2. **Assessment instruments:** A set of five principles and associated metrics to assess the quality (rigor and relevance) of canonical action research practice is proposed by Davison, Martinsons & Kock [185].

3. **KOS:** A literature survey identified 16 papers reporting action research in SE journals and classified them according to the study adherence, type, and control structures [396].

### 3.8.2.3 Design Science

<table>
<thead>
<tr>
<th>Planning + Execution + Analysis</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[240, 440]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Highlights

1. **Very few guidelines for SE research:** Although the studies cover the main process phases (including the evaluation of the design artifact) there is no specific support on how to report such studies.

### 3.8.2.4 Interviews

<table>
<thead>
<tr>
<th>Planning + Execution</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[344, 356, 374]</td>
<td></td>
<td></td>
<td>[318]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[213, 412]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[317, 398]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td></td>
<td>[362]</td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[243, 327, 391, 416]</td>
<td>[391]</td>
<td></td>
</tr>
</tbody>
</table>

#### Highlights

1. **Guidelines for interviews:** A comprehensive guide and a set of advice to plan and conduct semi-structured interviews is given in [243]. Its activities include:  
   * i) scheduling;  
   * ii) collecting of background information;
iii) preparing interview guides; iv) discussion/meetings; v) summary writing; and vi) transcribing. The guidelines aggregate existing knowledge from other disciplines and complement with authors’ own interviewing experiences.

- **Field studies:** Some papers address interviews as part of field study techniques, thus providing guidelines for conducting interviews along with observations and surveys [317, 356].

- **Ethical issues:** The application of interviews often face ethical dilemmas (e.g., consent, confidentiality, and beneficence). Singer and Vinson [416] identify ethical issues relevant to empirical studies, exemplified by real SE empirical studies.

- **Globally distributed:** Interviews can become more complex when interviewer and interviewee are not meeting face-to-face. Papers discussing such challenges and solutions are related to field studies [356] and qualitative methods [374, 412].

2. **KOS:** A taxonomy of field studies is given in the following [317, 318], discussing the advantages and disadvantages of each method (i.e., interviews are classified as direct and inquisitive, are also highly interactive but time-consuming).

### 3.8.2.5 Focus Group

<table>
<thead>
<tr>
<th>Planning + Execution + Analysis</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[311]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. **Guidelines for focus groups:** This method involves well-planned discussions to gather the perceptions of a group of participants [311]. The researcher acts as a mediator, facilitating and focusing the discussion on the topic.
3.8.2.6 Observation

<table>
<thead>
<tr>
<th>Planning</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Execution</td>
<td>[355,394]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>[215]</td>
<td>[215]</td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>[162]</td>
<td>[441]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[316,356]</td>
<td>[318]</td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td>[412]</td>
<td></td>
<td>[456]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[412]</td>
<td></td>
<td>[452]</td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>[235,317,398]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[201]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis + Reporting</td>
<td>[201]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[299,327,391,416]</td>
<td>[299,391]</td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for observational studies*: None of the identified studies presented a guide to observations alone. Most of the guidelines incorporate it to case study research e.g., [327,391,416].
   - *Reporting observational studies*: A set of guidelines for reporting the outcomes of observations in the form of lessons learned is provided by Budgen and Zhang [162]. Additional guidance is proposed for observations along case studies [391] and experiments [441].
   - *Qualitative data analysis*: It is often employed together with observations, such as grounded theory (e.g., [215,235]) and thematic analysis (e.g., [327]).
   - *Think-aloud protocol*: It is mainly used for testing user interfaces, in which the method was subject to an evaluation [347]. An alternative called dialog-based protocol can be used in pair programming tasks, as proposed and validated by Xu and Rajlich [454].

2. *Assessment instruments*: A checklist for assessing observation and experiment research reports is given by Wieringa et al. [441].

3. *KOS*: Field study techniques can be categorized according to the human intervention degree [318] (i.e., observation is classified as direct or inquisitive technique).
### 3.8.2.7 Ethnography

<table>
<thead>
<tr>
<th>Ass. Instruments</th>
<th>Guidelines</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[171,336]</td>
<td></td>
</tr>
<tr>
<td>Execution</td>
<td>[175, 272, 355, 373, 394]</td>
<td>[171]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[204]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[210,309,383,412]</td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis + Reporting</td>
<td>[388]</td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td>[386]</td>
<td></td>
</tr>
<tr>
<td>Analysis + Reporting</td>
<td>[436]</td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[343,387]</td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for ethnography*: An overview of ethnography studies in SE is given by Rönkkö [387] and some guidelines are provided in the following references [343,355,383].

### 3.8.2.8 Think-Aloud

<table>
<thead>
<tr>
<th>Ass. Instruments</th>
<th>Guidelines</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning + Execution</td>
<td>[347]</td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[454]</td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for think-aloud protocol*: It is mainly used for testing user interfaces, in which the method was subject to an evaluation [347]. An alternative called dialog-based protocol can be used in pair programming tasks, as proposed and validated by Xu and Rajlich [454].
3.8.2.9 Archival Research

### Highlights

1. **Guidelines for archival research**: We found no step-by-step guide supporting the process in SE research. The need for such guidelines is supported by several examples of SE-related archives that can be addressed by such studies (e.g., software repositories, project artifacts, and models).
   - **Analysis of work artifacts**: Several data collection techniques for archival research (e.g., electronic databases of work personal, tool logs, and documentation) are discussed by Lethbridge, Sim, and Singer [317,318]. Their advantages and disadvantages are discussed, and some advice on how to keep the records are given.
   - **Replicating data mining**: Studies based on data retrieved from development repositories are especially suitable for reproduction. Best practices, tools, and recommendations to replicate empirical studies on mining software repositories (MSR) are given by Gonzalez-Barahona and Robles [229] and Robles and Germán [384].
   - **Data collection for content analysis**: Archival research can provide a systematic approach to collect data for content analysis techniques. Papers such as [175,201,319] address such joint approach.

2. **Assessment instruments**: A method for assessing the reproducibility of studies based on data retrieved from development repositories is proposed by Gonzalez-Barahona and Robles [229].

3. **KOS**: Also archival research figure the taxonomy of field study techniques [318] (i.e., classified as indirect and unobtrusive).
3.8.2.10 Postmortem Review

<table>
<thead>
<tr>
<th>Phases</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[203]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[252]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for postmortem reviews*: Three different methods for carrying out such reviews are provided in [203]. Another work [252] provide lessons learned on aggregating different studies, including postmortem reviews, controlled experiments and case studies.

3.8.2.11 Systematic Literature Review

<table>
<thead>
<tr>
<th>Phases</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[432, 443, 459]</td>
<td></td>
<td>[432, 443]</td>
</tr>
<tr>
<td>Analysis</td>
<td>[176, 177, 198, 231, 381, 448]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>[158, 190, 438]</td>
<td>[156, 158]</td>
<td>[190]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[450, 458]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td></td>
<td></td>
<td>[456]</td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td></td>
<td></td>
<td>[402, 457]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[134, 168, 210, 258, 324, 377, 380]</td>
<td>[134, 238, 325, 377]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td></td>
<td>[183]</td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[322]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[145, 153, 155, 187, 246, 277, 294, 295, 312, 331, 349, 376, 378, 421, 427]</td>
<td>[128, 294, 376, 378]</td>
<td>[187]</td>
</tr>
</tbody>
</table>
3.8 Appendix

Highlights

1. **Guidelines for systematic literature reviews**: The most influential SLR guidelines in SE were proposed by Kitchenham and Charters [277, 294]. Another template based on SLR protocols developed in medical area is described by Biolchini et al. [145, 187, 331].

   - **Supporting the intrinsic activities**: Secondary studies follow a systematic process that requires several phases and discrete activities. Specific support for such phases include: i) search strategies e.g., [200, 202, 291, 423, 447, 458, 461]; ii) primary study selection e.g., [122, 360, 431]; iii) quality assessment e.g., [247, 291]; iv) data extraction e.g., [433]; v) data synthesis e.g., [176, 198, 231]; vi) reporting the review e.g., [381, 438, 448]; and vii) replication e.g., [190, 281].

   - **Evidence-based software engineering (EBSE)**: SE research might benefit from an approach for integrating the results from secondary studies with practitioners’ needs and values. Such approach is proposed by Kitchenham, Dybå and Jørgensen [295]. Some guidelines address the pedagogical aspects of using EBSE by students [258, 349, 377, 378].

   - **Tools to support SLRs**: such as SLuRp, StArt, SLR-Tool and SLR-TOOL are identified and further analyzed by Marshall, Brereton and Kitchenham [324]. A set of features that such tools should possess has been developed and used as the criteria to evaluate the candidate tools [325].

2. **Assessment instruments**: The systematic process employed in secondary studies require particular validation for some of its phases, such as: i) defining the research question e.g., [183]; ii) search strategies e.g., [129, 249, 282, 291, 458]; iii) quality assessment e.g., [290]; and iv) reporting the review e.g., [156, 158]. The particular characteristics and benefits of mapping studies have been assessed in a series of studies e.g., [161, 280, 292, 293].

3. **KOS**: A scientific research ontology provide support to describe the knowledge regarding SLRs studies [187].
3.8.2.12 Systematic Mapping

|----------|------------------|------------------|-----------|

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Guidelines [448]</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning + Execution</td>
<td>458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td></td>
<td></td>
<td>[456]</td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td></td>
<td></td>
<td>[402, 457]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[134, 361]</td>
<td>[134]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td></td>
<td>[183]</td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>292</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>Analysis + Reporting</td>
<td></td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[153, 246, 277, 294, 376, 421]</td>
<td>[294, 376]</td>
<td></td>
</tr>
</tbody>
</table>

Highlights

1. **Guidelines for systematic mapping studies**: Even though general guidance for secondary studies is given in Section 3.4.1, a comparative study investigated the differences between SLR and mapping studies. Based on the results, a specific set of guidelines for systematic maps is proposed by Petersen *et al.* [361]. Benefits and challenges of such studies have also been investigated by Kitchenham, Bugden & Brereton [292].

2. **Assessment instruments**: Some measures of quality (e.g., completeness, reliability, effectiveness) for systematic mapping studies were proposed and validated in the given papers [161, 280, 293].
3.8 Appendix

3.8.2.13 Survey

<table>
<thead>
<tr>
<th></th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>286, 297, 304, 367, 368</td>
<td></td>
<td>219</td>
</tr>
<tr>
<td>Execution</td>
<td>[373]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>[192]</td>
<td></td>
<td>305</td>
</tr>
<tr>
<td>Reporting</td>
<td>[366]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>174, 195, 256, 298, 303, 356, 374</td>
<td></td>
<td>298</td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>296, 412</td>
<td></td>
<td>130</td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>165</td>
<td></td>
<td>452</td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>170, 317</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td>362</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>169, 217, 218, 273, 299, 370, 416, 421</td>
<td>299</td>
<td>218, 370</td>
</tr>
</tbody>
</table>

Highlights

1. *Guidelines for survey research:* Two general guidelines provide support for the whole process. The principles of survey research series [108] is a starting point for novice researchers to understand the process and its main challenges. Another seven-stage framework [273] present similar guidance based on lessons learned and include additional resources, such as the activities breakdown and a cover letter. Furthermore, a set of recommendations to replicate survey-based research is given by Cater-Steel *et al.* [170].

- **Field studies:** Some papers discuss guidance for the survey process along with other field study methods e.g., [217, 317, 356, 416, 452].
- **Qualitative and quantitative surveys:** As mixed research, surveys usually collect both qualitative and quantitative data. However, several guidelines favor one type over another, such as [165, 373, 374, 412] emphasize the perspectives of the subjects, whereas [192, 296, 300, 301] address quantified data collection and analysis.
- **On-line surveys:** Recommendations for on-line self-administrated surveys are given by Punter *et al.* [375]. The paper covers issues such
as tracking responses in real-time, motivating respondents to return, and identifying drop-out questions i.e., where people stop filling out the questionnaire.

- **Sampling strategies:** A proper sampling strategy is a key to generalizing the findings from a survey questionnaire. Such issues are discussed in part five of principles of survey research series [286]. Moreover, there is a set of papers [193–195] investigating proposing improvements for large-scale sampling in surveys.

2. **Assessment instruments:** The survey questionnaire is often the target of evaluation regarding validity and reliability aspects [273, 298]. A comprehensive checklist provides support for identifying relevant threats to several research methods, including surveys [362].

3. **KOS:** A framework for classifying and examining survey research is proposed by Pinsonneault & Kraemer [187]. Some other categorization instruments cover several empirical research methods, distinguishing surveys’ particular characteristics [218, 402, 452].

### 3.8.2.14 Delphi Method

<table>
<thead>
<tr>
<th>Planning + Execution + Analysis</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[351]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. **Guidelines for delphi method:** Guidelines for selecting appropriate experts and ensure the validity of the study in IS are given by Okoli and Pawlowski [351].
### 3.8.2.15 Experiment

<table>
<thead>
<tr>
<th>Step</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[259, 260, 284, 304, 329, 336, 357, 364, 405, 428, 432, 443]</td>
<td>[262, 354]</td>
<td>[189, 219, 262, 274, 364, 432, 443]</td>
</tr>
<tr>
<td>Execution</td>
<td>[137, 138, 230, 237, 245, 270, 373, 407, 434]</td>
<td>[166, 270, 271, 289]</td>
<td>[245]</td>
</tr>
<tr>
<td>Reporting</td>
<td>[131, 167, 190, 191, 228, 248, 253, 255, 366, 379, 397, 408, 414, 444]</td>
<td>[276, 278, 279, 426, 441]</td>
<td>[131, 190, 191, 228, 422]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[125, 126, 144, 157, 193, 195, 214, 254, 303, 310, 345, 348, 404, 411, 418, 419]</td>
<td>[188, 411]</td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td>[392, 442]</td>
<td></td>
<td>[456]</td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td>[403, 406]</td>
<td>[173, 403]</td>
<td>[300, 402, 403, 457]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[133, 135, 136, 147, 154, 208, 216, 220, 224, 250, 264, 265, 283, 296, 301, 321, 340, 413, 420, 439, 446]</td>
<td>[130, 199, 446]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>[165]</td>
<td></td>
<td>[452]</td>
</tr>
<tr>
<td>Planning + Analysis + Reporting</td>
<td></td>
<td></td>
<td>[234]</td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[149, 181, 227, 328, 333, 339, 398, 425]</td>
<td>[149, 229, 425]</td>
<td></td>
</tr>
<tr>
<td>Execution + Reporting</td>
<td>[410]</td>
<td></td>
<td>[410]</td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td>[201, 207, 342, 451]</td>
<td>[261, 362]</td>
<td></td>
</tr>
<tr>
<td>Analysis + Reporting</td>
<td></td>
<td>[211, 268, 399]</td>
<td></td>
</tr>
</tbody>
</table>
CERSE - Catalog for empirical research in software engineering: a systematic mapping study

Highlights

1. **Guidelines for experiments**: A set of activities for performing experiment studies in SE are discussed in the following references: [133, 135, 154, 251, 365]. Additional guidance include meta-analysis techniques (e.g., [220, 332]) and replicating experiments (e.g., [320, 330]). Some guidelines focus on the pedagogical aspects of teaching or/and learning how to conduct experiment studies [175, 439].

   - **Human-based experiments**: SE often study human subjects i.e., practitioners, students, researchers. Several papers cover sampling and data collection strategies (e.g., [193, 373]), human subjects-related bias (e.g., [340]), ethical issues (e.g., [125, 416]), knowledge transfer (e.g., [404, 408]), and replication (e.g., [321]).

   - **Technology-based experiments**: Are also known as computational experiments and benchmarks i.e., studies comparing the performance of computer systems. A series of experience reports provide guidance in performing such studies, such as [147, 208, 313, 413]. Additional papers covers the reporting (e.g., [248]) and replication (e.g., [229]) aspects.

2. **Assessment instruments**: Several studies present evaluation of experimental research, addressing aspects such as continuous feedback [271], external and internal validity [166, 199, 233, 354], and experimental reporting [268, 441].

3. **KOS**: Categorization of experiments also cover different features, such as context [274], theory usage [234], subjects, tasks and environments [223, 245, 422], and type of experimental replications [131, 189, 228, 410].
3.8.2.16 Quasi-Experiment

<table>
<thead>
<tr>
<th></th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[432]</td>
<td>[262]</td>
<td>[262,432]</td>
</tr>
<tr>
<td>Analysis</td>
<td>[192]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[157,195]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[136,283]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>[165]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[169,226,269]</td>
<td>[226]</td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for quasi-experiments*: A.k.a. natural experiments are non-randomized experiments often conducted in a field setting. An overview of how such studies are designed and analyzed, the threats to validity and results are reported, is presented by Kampenes et al. [269]. Moreover, lessons learned on designing, conducting and analyzing data from quasi-experiments are reported in the following references [157,283].

3.8.2.17 Simulation

<table>
<thead>
<tr>
<th></th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>[443]</td>
<td></td>
<td>[219,443]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td></td>
<td>[188]</td>
<td></td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td>[352]</td>
<td></td>
<td>[457]</td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for simulation studies*: A secondary study characterizing the SBS initiatives in SE resulted in a set of guidelines [188]. The guidelines aim to improve such studies’ understandability, replicability, generalization and validity.

2. *Assessment instruments*: A framework is proposed to assess the acceptability of the components in a simulation study (e.g., parametric model, real-world and simulated data, simulation results, experiment specification, and methodologies or techniques in use) [352].
### 3.8.2.18 Thematic Analysis

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[210]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[178]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[327]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. **Guidelines for thematic analysis:** Our systematic process identified a proposal for thematic synthesis process in SE through a checklist [178]. The process and outcome associated with each step are described and illustrated with examples from the literature.

2. **KOS:** An overview of the data analysis methods and their related characteristics contributes to a better understanding of the challenges in synthesizing SE research [179].

### 3.8.2.19 Content Analysis

<table>
<thead>
<tr>
<th>Execution</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>[175]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[374]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[210]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. **Guidelines for content analysis:** Examples on the application of qualitative content analysis are presented by the following references [179, 319, 374].
3.8 Appendix

3.8.2.20 Grounded theory

<table>
<thead>
<tr>
<th></th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>[146, 215, 231]</td>
<td>[179, 215]</td>
<td>[179]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[244]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[121]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[235]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[120]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Guidelines for grounded theory:* A model including fifteen recommendations is presented by Adolph, Hall and Kruchten [121] based on their experience using classical (i.e., Glaserian) grounded theory in the SE context. It is intended to help researchers interpret the canons of grounded theory in a manner that is relevant to software engineers.

   - *Lessons learned:* Most of the identified papers are experience reports on conducting grounded theory e.g., [120, 121, 235, 244]. These studies discuss insights when conducting such research, in addition to examples of its application in SE research.

3.8.2.21 Hermeneutics

<table>
<thead>
<tr>
<th></th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[163, 309]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlights**

1. *Very few guidelines for hermeneutics:* We identified just two studies (i.e. [163, 309]) presenting an overview of the method. Both papers outline the application through interpretive studies on the information systems field.
### 3.8.2.22 Statistical Analysis

<table>
<thead>
<tr>
<th>Planning</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>[259,262]</td>
<td>[197]</td>
<td>[262]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[197, 222, 239, 275, 335, 389]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td>[442]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td>[402]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis + Reporting</td>
<td>[266]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[170,333]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis + Reporting</td>
<td>[261]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis + Reporting</td>
<td>[211]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[327]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Highlights

1. **Guidelines for statistical analysis:** Statistical analysis includes an extensive set of techniques, and just some of them are addressed by guidelines in the particular context of SE research. As an example, a 10-step framework for statistical prediction of software characteristics is proposed by Rosenberg [389].

   - **Descriptive statistics:** Visual techniques for statistical analysis in empirical SE studies are proposed by Garcia et al. [222]. A case study illustrates how such visualizations may improve interpretation of the experimental data.

   - **Inferential statistics:** Context and generalization problems potentially resulting from statistical hypothesis testing approaches are discussed by Jørgensen and Sjoberg [259]. The paper attempt to formulate how we should conduct empirical software engineering studies.

2. **Assessment instruments:** Statistical significance, statistical power, effect size and magnitude plays important roles in hypotheses testing and therefore should be carefully handled. Some papers investigate the assessment of such pivotal characteristics in ESE studies e.g., [157, 211, 239, 266, 335].
3.8 Appendix

<table>
<thead>
<tr>
<th>Planning</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>[192, 196–198, 205, 231, 239]</td>
<td>[196, 197]</td>
<td>[262]</td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td></td>
<td>[402]</td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[220]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execution + Analysis</td>
<td>[322, 333]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.8.2.23 Meta-Analysis

**Highlights**

1. *Meta-analysis*: Meta-analytic techniques applied to SE studies are illustrated by Hayes [239]. Moreover, Dieste *et al.* [198] provide a set of rules to help to choose the proper aggregation technique. Finally, for applying such techniques, a heterogeneity analysis of the studies’ results is required; Dieste *et al.* [196] validates how accurately the statistical methods detect heterogeneity.

### 3.8.2.24 Overall empirical research

<table>
<thead>
<tr>
<th>Planning</th>
<th>Guidelines</th>
<th>Ass. Instruments</th>
<th>KOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Execution</td>
<td>[237, 417]</td>
<td>[257]</td>
<td>[150, 257, 445]</td>
</tr>
<tr>
<td>Analysis</td>
<td>[215, 334]</td>
<td>[215]</td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>[132, 363, 371, 379]</td>
<td></td>
<td>[236, 241, 385, 424]</td>
</tr>
<tr>
<td>Planning + Execution</td>
<td>[125, 212, 232, 358, 411]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis</td>
<td></td>
<td>[411]</td>
<td></td>
</tr>
<tr>
<td>Planning + Reporting</td>
<td>[403]</td>
<td>[403]</td>
<td>[186, 402, 403, 457]</td>
</tr>
<tr>
<td>Planning + Execution + Analysis</td>
<td>[267]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Execution + Reporting</td>
<td>[165, 221]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning + Analysis + Reporting</td>
<td>[388]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Phases</td>
<td>[169, 299, 312, 421]</td>
<td>[299]</td>
<td></td>
</tr>
</tbody>
</table>
Highlights

1. **Guidelines for empirical research**: Some papers aim to help at identifying the appropriate methods and discuss the implications of this choice to the future research e.g., [186, 219, 237, 299, 449].
   
   - **Describing a context**: In order to decide whether a piece of evidence is helpful for a given situation, a complete and accurate description of the context which the study was conducted is needed. Papers supporting the description and assessment of the research context include the following references [212, 363].

2. **KOS**: Ontologies, taxonomies and glossaries help researchers in describe the research, as also to better communicate its findings. Some literature presenting such tools include [219, 241, 385, 456]. Moreover, Happel and Seedorf [236] present a framework to classify and apply ontologies.

3. **Assessment instruments**: A proper validity analysis and proposal of strategies to mitigate the impact of threats are essential to ensure the quality of empirical research. Some example of such literature is [215, 237]. Additionally, ethical issues of SE empirical research are discussed by the following [125, 232, 415].
### 3.8.3 Papers Covering multiple Research Methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[391]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[318]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[416]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[362]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[452]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[171]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[421]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[217]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[142]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[139]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[412]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[327]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[374]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[373]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[456]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[299]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[215]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[406]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[316]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[309]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[384]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[407]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[402]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[443]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[457]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[165]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[169]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[219]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[303]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Continued on next page*
CERSE - Catalog for empirical research in software engineering: a systematic mapping study

Continued from previous page

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[366]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[301]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[304]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[130]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[218]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[300]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[296]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[220]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[364]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[420]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[173]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[413]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[350]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[444]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[285]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[399]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[446]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[254]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[252]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[276]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[163]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[204]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[317]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[201]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[356]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[398]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[355]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[394]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[432]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[441]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix

- **Continued from previous page**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[428]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[336]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[235]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[175]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[210]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[348]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[176]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[229]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[319]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[360]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[182]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[202]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[448]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[183]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[225]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[282]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[447]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[458]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[200]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[429]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[246]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[461]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[153]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[249]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[376]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[277]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[423]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[129]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[294]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[172]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Continued on next page*
CERSE - Catalog for empirical research in software engineering: a systematic mapping study

- Continued from previous page

|-----------|-----------|-----------------|-----|------------|-----------------|----------------|-------------|-------------|-------------|-------------|-------------|------------------|------------------|-----|------------------|--------|---------------|------------|-----------------|-----------|------------------|---------------|-----------------|-------------|------------------|-------------|------------------|-------------|
3.8.4 CERSE References


3.8 Appendix


3.8 Appendix


CERSE - Catalog for empirical research in software engineering: a systematic mapping study


3.8 Appendix


CERSE - Catalog for empirical research in software engineering: a systematic mapping study


[282] Barbara Kitchenham, Pearl Brereton, Mark Turner, Mahmood Niazi, Stephen Linkman, Rialette Pretorius, and David Budgen. The impact of
3.8 Appendix

limited search procedures for systematic literature reviews a participant-
observer case study. In Proceedings of the 2009 3rd International Sym-
posium on Empirical Software Engineering and Measurement, pages 336–

[283] Barbara Kitchenham, David Budgen, Pearl Brereton, Mark Turner,
S Charters, and S Linkman. Large-scale software engineering questions-

[284] Barbara Kitchenham, John Fry, and Stephen Linkman. The case against
cross-over designs in software engineering. In Software Technology and
Engineering Practice, 2003. Eleventh Annual International Workshop on,

methodology for evaluating software engineering methods and tools. Com-

research: part 5: populations and samples. ACM SIGSOFT Software

[287] Barbara Kitchenham and Shari Lawrence Pfleeger. Principles of survey
research part 6: data analysis. ACM SIGSOFT Software Engineering

[288] Barbara Kitchenham, Lesley Pickard, and Shari Lawrence Pfleeger. Case

[289] Barbara Kitchenham, Dag IK Sjøberg, O Pearl Brereton, David Budgen,
Tore Dybå, Martin Höst, Dietmar Pfahl, and Per Runeson. Can we eval-
uate the quality of software engineering experiments? In Proceedings
of the 2010 ACM-IEEE International Symposium on Empirical Software

[290] Barbara A Kitchenham, O Pearl Brereton, David Budgen, and Zhi Li.
An evaluation of quality checklist proposals-a participant-observer case
study. In 13th International Conference on Evaluation and Assessment in

[291] Barbara A Kitchenham, Pearl Brereton, Mark Turner, Mahmood K Ni-
azi, Stephen Linkman, Rialette Pretorius, and David Budgen. Refining


3.8 Appendix


CERSE - Catalog for empirical research in software engineering: a systematic mapping study


CERSE - Catalog for empirical research in software engineering: a systematic mapping study


CERSE - Catalog for empirical research in software engineering: a systematic mapping study


3.8 Appendix


168 CERSE - Catalog for empirical research in software engineering: a systematic mapping study


3.8 Appendix


[441] Roel Wieringa, Nelly Condori-Fernandez, Maya Daneva, Bela Mutschler, and Oscar Pastor. Lessons learned from evaluating a checklist for reporting experimental and observational research. In *Proceedings of the
CERSE - Catalog for empirical research in software engineering: a systematic mapping study


3.8 Appendix


Chapter 4

Aligning the Views of Research Quality in Empirical Software Engineering

This chapter is based on the following paper:


4.1 Introduction

Research quality is a central concept in software engineering (SE) research. In the context of evidence-based software engineering, the quality plays an important role when aggregating evidence. Quality standards are a means for evaluating the research quality, and are often operationalized as checklists to assess quality (e.g., [472,478,496]).

In a similar way to software quality, not all qualities of research can be achieved to the same degree simultaneously. For example, a system that should be highly secure may not have a high degree of usability due to authentication mechanisms one has to go through. Hence, software engineers need to make trade-off decisions between the quality dimensions in order to prioritize them.
In turn, to consistently prioritize them, they also need to have a common understanding of the definition of quality and its dimensions.

Siegmund et al. [493] investigated how the view of researchers differ with respect to two quality dimensions, namely the external and internal validity of studies. As an example for trade-offs they highlighted: “There is an inherent trade-off in empirical research: Do we want observations that we can fully explain, but with limited generalizability, or do we want results that apply to a variety of circumstances, but where we cannot reliably explain underlying factors and relationships? Due to the options’ different objectives, we cannot choose both.” Hence, Siegmund et al. [493] elicited preferences within the community in a very simple manner, namely by asking whether there is a preference for maximizing or rather minimizing internal or external validity. A key finding was that opinions varied greatly among the participants. Siegmund et al. did not discover other studies focusing on eliciting trade-off preferences. Hence, research on understanding preferences concerning research quality is needed as this informs which methods are chosen and how studies are designed. The study by Sigmund et al. was focusing on the community level. That is, if the community sees a greater need for results that apply to a variety of circumstances, this influences the way the research is conducted; e.g. in the above example the preference would be given to studies with a high number of participants and contexts such as surveys.

The aim of this study is to conduct a first step towards aligning the views of research quality in software engineering. Our overarching main research question is: How applicable is the model proposed by Mårtensson et al. [483] in the context of empirical software engineering? This is achieved by the following contributions:

**Contribution 1:** Similar to Sigmund et al.’s study, the relative priorities of quality dimensions from the perspective of software engineering researchers are investigated. We complement Sigmund et al.’s study by considering a larger set of quality dimensions as defined by Mårtensson et al. [483].

**Contribution 2:** The model by Mårtensson et al., which is a model for research quality, is used as a basis to identify a common understanding of the definition of empirical software engineering (ESE) research quality and its dimensions. At the same time, Mårtensson et al. highlight that the model needs to be evaluated in different research fields.

Similar aims and contributions were pursued by Barney and Wohlin [464], though their focus was on the quality of specific software systems. They [464]
investigated the priorities of quality standards among different groups to understand how different groups and roles differ. They showed how groups are aligned or misaligned and at the same time discovered specific deficiencies with respect to the definition of qualities for the investigated quality model.

To achieve our research contributions two investigations were conducted. The first one was a case study of the Software Engineering Research and Education Lab Sweden (SERL Sweden) to study alignment on research group level. The second was a focus group study from the perspective of the research community, represented by the participants of the International Software Engineering Research Network (ISERN). Given the similarity of the investigations, the method by Barney and Wohlin [463] was adapted to achieve our research contributions.

The investigations resulted in a revised version of the conceptual research quality assessment model by Mårtensson et al. [483] specialized for ESE. We show the needs for adjusting the structure and definition of quality dimensions to the context of empirical software engineering research. This is an important step to adapt the conceptual model to the community's view, or the field it is intended to. For the revised model we also discuss an exemplary scenario via applying the revised model to characterizing a quality assessment instrument. This provides the potential users with insights on how to operationalize the model to a concrete scenario.

The remainder of the Chapter is structured as follows: Section 4.2 presents the related work. Sections 4.3 and 4.4 describe the research method and provides the results for the case study and focus group, respectively. We discuss the results in Section 4.5. Section 4.6 concludes the Chapter.

4.2 Background & Related Work

Evaluating research quality is a fundamental topic that has been discussed by researchers for a long time [485]. Such discussions lead to results that are formulated in terms of quality concepts aimed to assess research's methodological aspects. Examples of such concepts are reliability [477], practical relevance [466] and generalizability [481].

Furthermore, quality concepts are often addressed with regard to a particular methodology: validity and reliability are strongly related to a positivist approach, whereas the practical relevance is emphasized by interpretivist studies [469]. Some researchers argue on trade-offs needed, but others argue in favour of an integrated view, in which both concepts are relevant.
4.2.1 A Conceptual Model of Research Quality

Aligned with an integrated view, Mårtensson et al. [483] proposed a generic framework for classification of research quality, comprising 23 dimensions divided into four main areas: A) credible, B) contributory, C) communicable, and D) conforming. Details of each criterion are given in Appendix 4.8.1.

This framework is modelled as a hierarchical structure, in which related dimensions are grouped into subareas. For example, rigorous is defined as research that is contextual, internally valid and reliable. Figure 4.1 illustrates the hierarchical relationship between the dimensions, their main areas, and subareas.

Figure 4.1: A representation of the 23 dimensions proposed by Mårtensson et al. [483] organized according to the proposed hierarchy.
Further, this framework aims to represent research quality regardless of research field; however, its authors acknowledge the value in assessing its suitability within the context of different research fields. Motivated by such, the work described herein examines the use of such framework in the context of the research community we belong to - Empirical Software Engineering community.

We aim to identify which dimensions are considered more important for this community, and also within sub-groups in the same community. Explicitly, we ought to determine the level of alignment between researchers regarding the perceived importance of the quality dimensions.

### 4.2.2 Methods to Determine Alignment

We used two methods to identify the level of alignment between researchers. First, the SAAMSQ - Stakeholder Alignment Assessment Method for Software Quality [463] is a stepwise approach to elicit the importance assigned to different dimensions of a given quality model. It allowed us to determine which dimensions are subject to shared or conflicting views of importance. The method comprises 7 steps:

1. Select a case context;
2. Identify key stakeholders;
3. Select/tailor a quality model;
4. Develop a data collection instrument;
5. Conduct data collection;
6. Analyze the results; and
7. Present the results to participants.

Although focused on the quality of software products and services, SAAMSQ has the potential to identify views of quality in different contexts. Also, we recognize similarities between software quality standards described in [489] and the conceptual model in [483] that supports the application in the context of our study.

Secondly, focus group allowed us to closely interact with researchers, fostering a discussion on the perceived importance of the quality dimensions. Focus group research employs mediated discussions to gather qualitative data from a group of participants regarding the investigated topic [24]. In general, focus group produces an in-depth understanding of the investigated phenomenon, its causes and implications.
We used the methods complementary to each other. SAAMSQ was first applied in the internal case of SERL Sweden (Section 4.3), in which we were more familiar to the participants. Insights gathered from this first study led us to design a more thorough focus group (Section 4.4) to both determine the alignment and discuss the opinions of the participants regarding the importance of quality dimensions.

4.3 Internal Case Study

4.3.1 Method

Our research methodology is a single case study employing a survey questionnaire to collect quantitative data. Specifically, the participants assigned values of importance to dimensions describing different aspects of research quality. We followed the guidelines by Runeson and Höst for conducting and reporting case study research [491].

Context

This study was conducted in the Software Engineering Research and Education Lab Sweden (SERL Sweden), part of the Department of Software Engineering in the Blekinge Institute of Technology (BTH)\(^1\). Currently, close to 50 people work at SERL, representing 18 different nationalities. A large part of the team consists of research practitioners, i.e. 7 professors, 2 associate professor, 13 assistant professor, and 19 PhD students. At the time we conducted our study the team was a little different, which can present discrepancy with the results provided in Section 4.3.2. As an example, Postdoctoral researchers were promoted to assistant professors.

In line with BTH’s mission, SERL’s research is grounded in innovation and close collaboration with industry. It has partnership with several companies that develop software intense systems, services, and products. Moreover, BTH is well-ranked academic research in empirical and evidence-based software engineering [476].

The members of SERL Sweden are also acknowledged for providing methodological guidelines for researchers, covering a wide range of methods such as experiments (e.g. [497]), case studies (e.g. [488]), systematic literature and mapping studies (e.g. [474,486]), and mixed methods studies (e.g. [471]).

\(^1\)https://www.bth.se/eng/about-bth/organisation/faculty-of-computing/dipt/
Research questions and proposition

This case study’s goal was to investigate the relative importance of the quality dimensions, as per the previously presented framework, based upon participants’ judgements. Two research questions (targeting Contribution 1 and 2 presented in Section 4.1) were formulated as means towards achieving such goal, as follows:

- **RQ1a** What are the most important research quality dimensions in SE research, as per the participants?

- **RQ1b** What is the level of alignment between the different participants with regard to their assessment of the importance of quality dimensions for SE research?

Our proposition for this research is that an alignment of the perceived importance of research quality represents the level of agreement within the context of this group. To recall, the study by Mårtensson [483] found very different preferences, which may originate from the spread of participants between different program committees with different foci.

Subjects

Subjects were selected through convenience sampling from the researcher positions at SERL, i.e. full professors, associate professors, assistant professors, and Ph.D. students. We sent invitations to the candidates and clustered the ones willing to participate in groups up to five.

Unit of analysis

The unit of analysis is the relative importance assigned to each dimension of research quality. Relative importance may show distances between the interpretations and permits to demonstrate which are the most critical dimensions in comparison to others.

We designed a data collection instrument to allow for subjects to express to what extent the dimensions are important. Subjects assigned importance values using hierarchical cumulative voting [490]. First, they should distribute 1000 points between the four main areas (Credible, Contributory, Communicable, and Conforming) in the Framework, and later distribute another 1000 between the subcriterias related to each area. In this way, the values for each subcriteria are weighted according to the values assigned to its area. Here the higher the number of points, the higher the importance (and hence priority).
Cumulative voting encourages subjects to prioritize and make trade-offs, given that whenever they add points to one criterion, they have to also remove points from another criterion, so to keep the total intact. This way it becomes more difficult to simply give the same number of points to every criterion.

Data collection

We conducted four workshops to gather data from 17 participants regarding the importance of the Framework’s quality dimensions. Due to convenience, all workshops took place in a meeting room in the work environment. They occurred during the same week, and sessions lasted from 40 to 60 minutes. We asked the participants to not shared details of the study to others outside of their workshop sessions.

At the beginning of each session, the first author acting as a moderator informed the participants about the aim of the study and the structure of the workshop. They were also asked for their consent, prior to gathering any data. Further, the moderator presented the Framework, followed by a few minutes for any questions/doubts to be asked by the participants.

Participants were instructed to bring their laptops, as we provided them with an electronic spreadsheet\(^2\), which gathered data on their assessment, and also some demographics data.

With the moderator being present, each participant filled the spreadsheet individually. They were allowed to ask any further clarification and even to express their overall ideas to the colleagues in the same workshop session. We took notes of the participants’ interaction with the moderator and among themselves. Next, the moderator explained the structure of the instrument and provided any further clarifications, as per the participants’ needs.

Participants filled out the electronic spreadsheet individually. They were also allowed to ask the moderator for clarifications and to express their overall understanding of a given criterion to the colleagues in the same workshop session. We took notes of the participants’ interaction with the moderator and between themselves.

Data analysis

To facilitate answering our research questions, we aggregated participants’ answers into a shared data frame\(^3\). We further normalized the values by weighting

\(^2\)available at: https://bit.ly/2InnyhG
\(^3\)available at: https://bit.ly/2InmXMy
4.3 Internal Case Study

the points assigned to each subcriterion taking into account the points allocated to its related main area. The equation for normalized importance is: 
\[(a \times n(a) \times c^a) / 1000\], where:

- \(a\) is the amount of points assigned to a main area,
- \(n(a)\) is the number of subcriteria related to main area \(a\), and
- \(c^a\) is the amount of points associated to a particular subcriterion.

As an example, given that a participant assigned 225 points to main area A, which contains six subcriteria; and 150 points to the subcriteria A1, the resulting normalized value for A1 is \((225 \times 6 \times 150) / 1000 = 202.5\).

The normalization standardizes the range of values over all the subcriteria, while at the same time considering the relative importance of the related main area. It also ensures that the probability of assigning values by chance is the same for all subcriteria. A similar procedure is employed by [480] for hierarchical voting analysis.

After computation, a given participant’s relative importance values always adds up to 1000 points. This allows for the comparison between the relative importance assigned by different participants or participants’ subgroups (e.g. PhD students).

Further, we provide an overview of the values assigned to each criterion from all participants through boxplots. This method for graphical representation of distributions gives a good indication of the dispersion of the data. Less dispersed distributions of importance values represent a higher degree of alignment among participants.

To interpret the results of the boxplots, we computed an importance threshold of 250, i.e. sum of all points \((t = 1000)\) divided by the total number of main criteria \((n = 4)\). Similarly to Rovegård, Angelis & Wohlin [490], we drew four distinct importance categories based on the characteristics of the distribution in relation to the threshold:

- **Unimportant (U):** the third-quartile is lower than the importance threshold.
- **Lower importance (LI):** importance threshold is between the median value and the third-quartile.
- **Important (I):** importance threshold is between the first-quartile and the median value; and
- **High importance (HI):** the first-quartile is higher than the importance threshold.
Finally, we calculated a priority list by ranking the quality dimensions according to the median value of the importance distributions. Such priority list represents the quality aspects most important, based upon participants’ assessments.

Validity threats

Next we discuss a series of potential validity threats, using as basis the four aspects of the validity for case study research described by Runeson and Höst [491]. We also detail the procedures taken to handle the identified threats.

**Construct validity:** One potential threat relates to results bias if the data collection instrument is not designed carefully. To mitigate such threat, we opted to use the hierarchical voting prioritization approach, which has been successfully applied in several studies (e.g. [480,490]). These studies’ reported experiences contributed towards the creation of our questionnaire. Prior to using the instrument, we tested it in a pilot with the participation of six subjects presenting similar research profile to our participants.

The quality dimensions are grounded on the work of Mårtensson et al. [483], which has been validated in a multidisciplinary context. To mitigate any misunderstandings relating to the framework, we presented and discussed the Framework and its elements in the beginning of the workshops.

**Internal validity:** Our case study aims to identify the participants’ view of the quality dimensions for overall SE research. However, there is a limitation associated with the research preferences of each participant. For example, a researcher that conducts mainly experiment studies is likely to interpret the quality framework with a more positivist view. To mitigate the threat related to a narrow spectrum of SE research, we instructed the participants to consider the dimensions on a broader ESE context.

**External validity:** Our study was conducted in the particular context of the research practitioners of SERL Sweden. The candidates were selected based on convenience and produced a quite small sample. From the 17 participants, 9 are PhD students, three Postdocs, three Assistant professors and only one Professor. As there are more junior than senior researchers, we could not assume that our participants represent the overall population of researchers in software engineering.

Our sample is reasonably heterogeneous with regard to their familiarity to different research methodologies and topics of investigation (see Section 4.3.2). Most of them has also experience in collaboration with industry, but only one
is familiar to pure theoretical research. Such characteristics of the sample limit our discussion to the context of the observations.

Reliability: There is a potential threat related to the misinterpretation of collected data. To mitigate such threat, both data analysis and interpretation procedures were designed in advance, based on the works of Kuzniarz & Angelis [480], and Petersen, Khurum & Angelis [487]. Mainly, we normalized the importance values, prior to interpretation, given the 23 dimensions were not equally distributed across the four main areas. All the transformations applied over the data are provided in our data set\(^2\) so that other researchers can validate and replicate our results.

4.3.2 Results

Participants’ Demographics

In total, 17 researchers participated in our case study, from which 9 are PhD students, three Postdocs, three Assistant professors and only one Professor. In average, they have been performing (and also reviewing and assessing) research in SE for more than 7.5 years, with a maximum of 30 years and a minimum of least than a year.

For the research type, most of the participants conduct empirical research in collaboration between academia and industry (9 out of 17 participants) or mostly in industry setting (3 participants). Just one researcher mentioned being familiar with pure theoretical research. Case study is the most prolific research method employed (by 12 participants), followed by systematic literature review (9), systematic mapping (6), interviews (6), survey-based research (5) and quasi-experiments (5). Other research methods the participants are familiar with include action research, archival research, design science, experiment, focus group, meta-analysis, observations, simulation, and thematic analysis.

Their research covers a wide range of SE topics, such as Software Testing (7 participants), SE Management (3), SE Process (3), Requirements Engineering (3), Software Construction (3) among others. The participants have published an average of 7.5 journals and 13.8 conference papers. The preferred venues for publication of their work are IST - Information and Software Technology (13 participants), JSS - Journal of Systems and Software (11), and EMSE - Empirical Software Engineering (6). The preferred conferences are ESEM - Empirical Software Engineering and Measurement (6 participants) and EASE - Evaluation and Assessment in Software Engineering (4).
RQ1a) Importance of the Quality Dimensions

Boxplots showing the distribution of importance assigned to all quality dimensions are presented in Figure 4.2 (four main areas on top and subcriteria on the bottom). The colours used in all the boxplots relate to the four main areas, as follows: blue for credible (A), red for contributory (B), green for communicable (C) and yellow for conforming (D). Boxplots are displayed according to the priority list, i.e., higher medians are shown on the top, and lower medians at the bottom. Further, a threshold line is drawn vertically over the plot, thus allowing us to divide it according to importance categories, from High (HI) at the top, to Low (U), at the bottom.

The main areas credible (A) and contributory (B) are ranked as more important than the other two main areas, with a median above the threshold; followed by communicable (C), and finally conforming (D). With regard to the main areas’ subcriteria, except for A3, all subcriteria for the main area credible (A) have medians above the threshold level, thus clearly suggesting a higher level of importance. Most of the subcriteria for contributory (B) also have medians above the threshold level, when compared to the subcriteria for the other two main areas communicable (C) and conforming (D). The subcriteria C1.2, D2.1 and C1.3 are the only three subcriteria from the main areas (C) and (D) that have medians at or above the threshold, respectively.

Five dimensions were considered of higher importance to the participants: B2.2 applicable results, B2.1 relevant research idea, A1.2 reliable, B1.3 original result, and A1.1 internally valid, respectively. It is important to notice that this set comprises 2 out of 3 dimensions related to the rigorous and relevant aspects. Lower importance dimensions are mostly associated with communicable (C) and conforming (D) main areas. In particular, the unimportant (U) subcriteria are related to these two areas. The only subcriteria associated to the communicable (C) rated important is C1.2 understandable and C1.3 readable. From the conforming (D) area, D1.1 morally justifiable is the only subcriterion rated important.

RQ1b) Participants’ Alignment

The characteristics of the distributions of importance in Figure 4.2 are also indicators of the level of alignment between participants. Shorter boxes and tail length denote a higher level of agreement over the number of points given, whereas larger boxes would be the opposite.
4.3 Internal Case Study

Figure 4.2: Distribution of importance values assigned to the four main areas (top) and relative importance values computed to the subcriteria (bottom). Relative importance values for the subcriteria are normalized by weighting the points assigned to each subcriterion by the points allocated to its related main area.
Besides the visual representation, characteristics of the distribution are useful to identify levels of alignment. In particular, measures of variation, such as the standard deviation, quantifies the dispersion of a distribution. A low standard deviation means that the assigned values of importance are concentrated close to an expected value, thus denoting alignment. Table A1.1 summarizes the characteristics of the distributions that could be used to determine alignment.

Table A1.1: Summary of the characteristics of the distributions for the subcriteria presented in Figure 4.2 (bottom). Each subcriterion is presented in a row, followed by its importance category; minimum, median and maximum values for the distribution; and the standard deviation ($\sigma$).

<table>
<thead>
<tr>
<th>Sub-criterion</th>
<th>Category</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>SD ($\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1.1 Internally valid</td>
<td>HI</td>
<td>150</td>
<td>288.8</td>
<td>1200</td>
<td>254.09</td>
</tr>
<tr>
<td>A1.2 Reliable</td>
<td>HI</td>
<td>225</td>
<td>315</td>
<td>900</td>
<td>192.51</td>
</tr>
<tr>
<td>A1.3 Contextual</td>
<td>I</td>
<td>120</td>
<td>265.2</td>
<td>450</td>
<td>76.73</td>
</tr>
<tr>
<td>A2 Consistent</td>
<td>I</td>
<td>165</td>
<td>265.2</td>
<td>480</td>
<td>91.07</td>
</tr>
<tr>
<td>A3 Coherent</td>
<td>LI</td>
<td>120</td>
<td>240</td>
<td>275.6</td>
<td>51.55</td>
</tr>
<tr>
<td>A4 Transparent</td>
<td>I</td>
<td>135</td>
<td>250</td>
<td>480</td>
<td>82.33</td>
</tr>
<tr>
<td>B1.1 Original Idea</td>
<td>I</td>
<td>0</td>
<td>283.5</td>
<td>840</td>
<td>190.46</td>
</tr>
<tr>
<td>B1.2 Original Procedure</td>
<td>I</td>
<td>0</td>
<td>250</td>
<td>700</td>
<td>160.21</td>
</tr>
<tr>
<td>B1.3 Original Result</td>
<td>HI</td>
<td>70</td>
<td>308</td>
<td>875</td>
<td>187.06</td>
</tr>
<tr>
<td>B2.1 Relevant Research Idea</td>
<td>HI</td>
<td>0</td>
<td>315</td>
<td>525</td>
<td>114.63</td>
</tr>
<tr>
<td>B2.2 Applicable Result</td>
<td>HI</td>
<td>105</td>
<td>346.5</td>
<td>1120</td>
<td>218.03</td>
</tr>
<tr>
<td>B2.3 Current Idea</td>
<td>LI</td>
<td>0</td>
<td>231</td>
<td>364</td>
<td>106.58</td>
</tr>
<tr>
<td>B3 Generalizable</td>
<td>LI</td>
<td>0</td>
<td>175</td>
<td>560</td>
<td>148.47</td>
</tr>
<tr>
<td>C1.1 Structured</td>
<td>U</td>
<td>0</td>
<td>168.8</td>
<td>500</td>
<td>115.40</td>
</tr>
<tr>
<td>C1.2 Understandable</td>
<td>I</td>
<td>156.2</td>
<td>281.2</td>
<td>750</td>
<td>144.24</td>
</tr>
<tr>
<td>C1.3 Readable</td>
<td>I</td>
<td>0</td>
<td>250</td>
<td>500</td>
<td>112.13</td>
</tr>
<tr>
<td>C2 Accessible</td>
<td>LI</td>
<td>0</td>
<td>187.5</td>
<td>343.8</td>
<td>89.83</td>
</tr>
<tr>
<td>C3 Searchable</td>
<td>LI</td>
<td>0</td>
<td>187.5</td>
<td>275</td>
<td>76.42</td>
</tr>
<tr>
<td>D1 Aligned with Regulations</td>
<td>LI</td>
<td>50</td>
<td>190</td>
<td>312.5</td>
<td>74.49</td>
</tr>
<tr>
<td>D2.1 Morally Justifiable</td>
<td>I</td>
<td>100</td>
<td>253.5</td>
<td>328.1</td>
<td>68.98</td>
</tr>
<tr>
<td>D2.2 Open</td>
<td>LI</td>
<td>50</td>
<td>240.6</td>
<td>328.1</td>
<td>91.80</td>
</tr>
<tr>
<td>D2.3 Equal Opportunities</td>
<td>LI</td>
<td>0</td>
<td>200</td>
<td>312.5</td>
<td>89.97</td>
</tr>
<tr>
<td>D3 Sustainable</td>
<td>U</td>
<td>0</td>
<td>112.5</td>
<td>500</td>
<td>126.79</td>
</tr>
</tbody>
</table>
A close alignment is represented in the boxplot by smaller boxes and short whiskers, e.g. A1.3 contextual ($\sigma = 37.36$) and A3 coherence ($\sigma = 44.48$). They are consistently ranked (important and lower importance, respectively) across the quality dimensions. This implies that the views towards the importance of these criteria are more aligned and do not change much within the participant group.

As examples of misalignment, the subcriteria A1.1 and A1.2 ($\sigma = 254.09$ and 192.51, respectively) are widely distributed, thus suggesting that the participants do not share a common perception about the importance of these dimensions. They are also skewed to the right, indicating that the assigned importance tends to the lowest values. Similar misalignments exist with relation to several contributory subcriteria, e.g., B2.2 applicable result ($\sigma = 218.03$) and B1.1 original idea ($\sigma = 190.46$). The distribution of importance ratings for these subcriteria is highly influenced by the outliers.

Outliers are represented by small circles beyond the bounds of the whiskers. Herein they point our importance rates assigned by a participant that does not agree with most. The outliers are very common in our boxplot of the normalized importance values (illustrated in Figure 4.2), except for four dimensions (D2.1, D2.2, B2.3, and C3, respectively). It is interesting to note that outliers are more common towards higher values of importance.

Several distributions have a lower limit equals to zero (0). Explicitly, several subcriteria on the lower importance (LI) and unimportant (U) categories fit this description. This means that the criterion is considered non-important (assigned value = 0) by at least one participant. In particular, B3 generalizable, C2 accessible and D3 sustainable are assigned a non-important for more than one participant. About B3, one of the participants commented: "[it] depends on the type of the study, e.g., case study research is not generalizable at all outside the context."

4.4 Focus Group

4.4.1 Method

We employed a focus groups methodology to gather experts’ opinion regarding the importance of research quality dimensions, where such dimensions are also based upon the model of research quality by Mårtensson et al. [483]. Further, we summarize our research method according to the guidelines provided by Kontio, Lehtola & Bragge [479].
Defining the research problem

Our second study aimed to bring together the perceptions of senior ESE researchers towards the dimensions of research quality as per the above mentioned framework. Besides the relative importance of the dimensions, we also wanted to promote discussions and knowledge building on what constitutes “research quality”. The following research questions geared this study:

RQ2a What are the most important research quality dimensions in SE research, as per the participants?

RQ2b What is the level of alignment between the different participants with regard to their assessment of the importance of quality dimensions for SE research?

RQ2c To what extent the model of research quality proposed by Mårtensson et al. [483] is relevant for software engineering research?

Research questions RQ2a and RQ2b address Contribution 1 while RQ2c addresses Contribution 2 (see Section 4.1).

Selecting the participants

The study was conducted within the context of the International Software Engineering Research Network (ISERN), which is a community of SE researchers (in particular senior/experts) and practitioners who apply an evidence-based paradigm to SE research. These SE researchers also have knowledge on methodological research, and commonly employ (and some also propose) different research methods to conduct empirical studies in SE.

Further, this community also contributes with steering the research community towards the quality of research, and discussions on the subject have led to publications that raise awareness of the importance of the quality and appropriateness of the methods used in the SE empirical studies, e.g. [462,465,475,494].

ISERN holds annual meetings as part of the Empirical Software Engineering International Week (ESEIW). Thus, we proposed the ISERN organization to conduct a focus group study during their meeting. Further, we sent email invitations to ISERN participants and also advertised the study during the ESEIW.
Planning and conducting the focus group session

We designed the focus group to fit a session lasting 1.5 hours, consisting of six parts. The first author acted as mediator, fostering discussion among the participants, while the second author served as an observer, taking notes and recording additional data. The participants were allowed to ask questions to the moderator and to the other co-participants during the whole session.

We conducted two sessions, the first one with eight participants and the second one with four participants. Each session started with an introduction in which we provided an overview of our objective, i.e., to discuss what constitutes “research quality” according to their perspectives. We also provided practical instructions and asked participants for verbal consent for audio recording.

Further, we presented the conceptual model of research quality [483], describing its components and structure. This model was used as a starting point for the discussion, and we carried out a prioritization exercise similar to the one done in our case study (see Section 4.3.1). Next, we provided participants with a whiteboard in which they could share their main opinions visually. We asked them to use green and red post-its to point out the most and the least important subcriteria for each of the four main areas. This representation was used to guide the discussions in the last part of the session.

Finally, we moderated a semi-structured discussion covering all the dimensions of the conceptual model. In particular, we focused the discussions on potential misalignments, i.e., whether dimensions were assigned high (green post-its) and low (red post-its) importance by different participants.

Analysis

The focus groups sessions were documented using four different media: 1) filled out prioritization questionnaires, 2) whiteboard containing the visual representation of the highest and lowest important dimensions, 3) audio recordings of the sessions and their transcripts, and 4) notes from the observer.

Out of 12 participants, 10 filled out the prioritization questionnaires. To analyse the importance of the quality dimensions, we employed a hierarchical voting analysis similar to the one used in our case study (see Section 4.3.1). The prioritization values of all participants were aggregated into a shared data frame. Later, we computed the relative importance values, and the subcriteria with the values assigned to the four main dimensions.

---

4 Script for focus group sessions is available at: https://bit.ly/2REKUmC
5 available at: https://bit.ly/2BGHpTr
To interpret the results, we used the same importance threshold and categories described in Section 4.3.1. The normalized values of importance and distributions are therefore comparable to the results of our case study. We also used the visual representation of the post-it board to triangulate the prioritization results.

The transcripts of the focus group discussions were later analysed by the first and second authors, using pattern-matching and vote counting [479]. The opinions of the participants were summarized into a short sentence and then mapped to related dimensions. Further, we counted the participants who mentioned or agreed with such an idea. We also identified cases in which conflicting ideas were provided, e.g. whether to merge A4 Transparent to another dimension or to keep it separate.

The analysis resulted in the identification and categorization of opinions with regard to the dimensions of quality. It also provided insights about improvement needs of the conceptual model with regard to dimensions that are redundant, dependent or poorly described. Such suggestions were also organized and used to revise the model.

We also used the observer’s notes to cross-check participants’ opinions, as per the moderator’s recollection. Moreover, the post-it board allowed us to confirm whether the comments from participants were related to a divergent opinion regarding a quality dimension. This further triangulation was used to support the decision making regarding the revision of the conceptual model.

**Validity Threats**

The potential threats to the validity of our focus group results are discussed with regard to four categories, as suggested by Runeson and Höst [491]:

**Construct validity:** The Framework of research quality has been created and evolved during workshops with experts from multiple research fields [483]. To minimize any potential threat related to the incorrect interpretation of the quality dimensions, we presented and discussed the Framework’s structure and its components during the focus group session.

Further, the participants could also discuss between themselves, so aiming to have a common understanding of the conceptual model’s dimensions and sub-criteria. The moderator also suggested participants share conflicting opinions, so seeking a complete interpretation of the dimensions through the divergent thinking process.

**Internal validity:** To ensure the trustworthiness of our study, the first author designed a focus group protocol, later evolved by discussing it with the
second author. Finally, the third and fourth authors reviewed the design by reflecting on the understandability of the questions and the accuracy of data collection. To reduce researcher bias, during the focus group session, the observer also acted as moderator’s referee, keeping a record of any action that could threaten the nature of the discussion. None of such actions were identified.

In general it is difficult to keep balanced group dynamics in focus groups sessions. We employed a semi-structured design to ensure that all the participants had the opportunity to express their opinion. Later on, we assessed the number of suggestions provided by each participant, to identify the ones that influenced the activity more. Participants were self-selected from the ISERN members attending their annual meeting.

**External validity:** Our focus group gathered opinions from a small group of experts who are senior researchers and have the expertise to judge the topic of the discussion, i.e. quality of research in software engineering. They are also members of a leading community in software engineering research.

To become an ISERN member, one has to apply (with their institution) and evaluated by a senior committee who decided whether their application will succeed or not. Although our sample does not represent the diversity of the SE research population, it supports the transferability of our findings by expertise, i.e. our results are drawn from the opinions of experts who are likely to influence the broader community.

**Reliability:** The discussions and interaction of the participants provided us with rich qualitative data reflecting their opinions. To ensure that the opinions were correctly understood by the researchers, the moderator often paraphrased the participants to obtain confirmation. Later, the analysis and interpretation of the qualitative data were carried out according to the study design. The moderator summarized and categorized the opinions, further reviewed by the observer.

### 4.4.2 Results

**Participants’ Demographics**

Twelve members from the ISERN participated in our focus group study. Most of them are academics: professors (5 participants), associate professor, assistant professor, and post-doctoral researcher (1 participant each). Two participants work in other research positions, i.e. research scientist and head of research department. Finally, one of the participants works on the industry as IT consultant.
The participants have published an average of 17.4 journals and 63.5 conference papers, well above the participants in the case study. Just five of them provided us detailed information regarding their research preferences and previous experiences. Research methods mentioned by these participants include case study research, design science, interviews, systematic literature review, systematic mapping, experiments, quasi-experiment, statistical and qualitative analysis.

**RQ2a) Importance of the Quality Dimensions**

Here boxplots were also used to display the distribution of values (level of importance) assigned to each quality dimension (see Figure 4.3 top and bottom). We ordered the subcriteria by the median values and used the threshold (vertical red line) to classify them according to importance. The boxes’ colours represent the main areas to which the subcriteria are grouped, i.e. blue for credible (A), red for contributory (B), green for communicable (C) and yellow for conforming (D).

One can notice some similarities with our case study’s results, such as the overall importance assigned to the main area (A), followed by (B), and then (C) and (D). Despite that, the data shows that the ISERN experts favour the subcriteria related to credible (A) main area over all the others, except for A3 Coherent, ranked as lower importance. Two subcriteria (A1.1 and A4, respectively) are considered of high importance (HI) according to our categorization, as their boxes are placed above the threshold.

With regarding the contributory (B) aspect, three subcriteria (B3, B2.2, and B2.1) have a median above the threshold, thus categorized as important (I). The four other subcriteria (i.e., B1.3, B2.3, B1.1, and B1.2) have medians below the threshold. Specifically, the third quartile of B2.3 current idea is below the threshold, thus classified unimportant (U) according to our categorization. All the subcriteria related to communicable (C) and conforming (D) areas have medians below the threshold. It worth mentioning that the values are obtained by HCV. Thus the values assigned to the main areas impacts the final result for each related subcriteria.

These boxplots aggregate and summarize all participants’ assessments, without being specific about any individual assessment. Exceptions to this are the outliers, i.e. data points that are abnormally far from all other participants. The outliers are less frequent than our case study, but it is important to note that the sample size of the focus group is smaller (only 10 participants) and thus likely to impact on the distribution of quantitative data.
Figure 4.3: Distribution of importance values assigned to the four main areas (top) and relative importance values computed to the subcriteria (bottom). Relative importance values for the subcriteria are normalized by weighting the points assigned to each subcriterion by the points allocated to its related main area.
The extent of the bars and the whiskers suggest conflicting opinions among the participants. The conflict is also pointed out by the standard deviation, as presented in Table A1.2. A higher deviation means that the relative importance values are dispersed, thus suggesting misalignment. This is the case for B3 generalizable (\(\sigma = 397.38\)) and B2.1 relevant research idea (\(\sigma = 325.19\)), which are ranked as the most important for some participants and as of low importance by others. This potential misalignment was the focus of our discussions during the focus group and provided us deeper insights regarding the participants’ opinions.

Table A1.2: Summary of the characteristics of the distributions for the subcriteria presented in Figure 4.3 (bottom). Each subcriterion is presented in a row, followed by its importance category; minimum, median and maximum values for the distribution; and the standard deviation (\(\sigma\)).

<table>
<thead>
<tr>
<th>Sub-criterion</th>
<th>Category</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>SD ((\sigma))</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1.1 Internally valid</td>
<td>HI</td>
<td>270</td>
<td>435</td>
<td>630</td>
<td>116.32</td>
</tr>
<tr>
<td>A1.2 Reliable</td>
<td>I</td>
<td>105</td>
<td>315</td>
<td>480</td>
<td>131.09</td>
</tr>
<tr>
<td>A1.3 Contextual</td>
<td>I</td>
<td>108</td>
<td>300</td>
<td>500</td>
<td>124.34</td>
</tr>
<tr>
<td>A2 Consistent</td>
<td>I</td>
<td>0</td>
<td>315</td>
<td>500</td>
<td>153.97</td>
</tr>
<tr>
<td>A3 Coherent</td>
<td>LI</td>
<td>150</td>
<td>225</td>
<td>500</td>
<td>126.87</td>
</tr>
<tr>
<td>A4 Transparent</td>
<td>HI</td>
<td>150</td>
<td>375</td>
<td>900</td>
<td>208.61</td>
</tr>
<tr>
<td>B1.1 Original Idea</td>
<td>LI</td>
<td>52.5</td>
<td>183.8</td>
<td>300</td>
<td>86.38</td>
</tr>
<tr>
<td>B1.2 Original Procedure</td>
<td>LI</td>
<td>0</td>
<td>147</td>
<td>300</td>
<td>126.35</td>
</tr>
<tr>
<td>B1.3 Original Result</td>
<td>LI</td>
<td>52.5</td>
<td>210</td>
<td>577.5</td>
<td>150.85</td>
</tr>
<tr>
<td>B2.1 Relevant Research Idea</td>
<td>I</td>
<td>140</td>
<td>281.2</td>
<td>1120</td>
<td>325.19</td>
</tr>
<tr>
<td>B2.2 Applicable Result</td>
<td>I</td>
<td>140</td>
<td>290</td>
<td>630</td>
<td>167.55</td>
</tr>
<tr>
<td>B2.3 Current Idea</td>
<td>U</td>
<td>35</td>
<td>183.8</td>
<td>300</td>
<td>94.17</td>
</tr>
<tr>
<td>B3 Generalizable</td>
<td>I</td>
<td>105</td>
<td>290</td>
<td>1470</td>
<td>397.38</td>
</tr>
<tr>
<td>C1.1 Structured</td>
<td>U</td>
<td>50</td>
<td>200</td>
<td>350</td>
<td>90.15</td>
</tr>
<tr>
<td>C1.2 Understandable</td>
<td>LI</td>
<td>75</td>
<td>200</td>
<td>375</td>
<td>91.52</td>
</tr>
<tr>
<td>C1.3 Readable</td>
<td>U</td>
<td>0</td>
<td>200</td>
<td>300</td>
<td>93.92</td>
</tr>
<tr>
<td>C2 Accessible</td>
<td>LI</td>
<td>93.75</td>
<td>250</td>
<td>450</td>
<td>110.79</td>
</tr>
<tr>
<td>C3 Searchable</td>
<td>LI</td>
<td>0</td>
<td>150</td>
<td>450</td>
<td>143.95</td>
</tr>
<tr>
<td>D1 Aligned with Regulations</td>
<td>LI</td>
<td>25</td>
<td>200</td>
<td>300</td>
<td>89.22</td>
</tr>
<tr>
<td>D2.1 Morally Justifiable</td>
<td>LI</td>
<td>100</td>
<td>200</td>
<td>600</td>
<td>139.54</td>
</tr>
<tr>
<td>D2.2 Open</td>
<td>LI</td>
<td>50</td>
<td>200</td>
<td>625</td>
<td>157.37</td>
</tr>
<tr>
<td>D2.3 Equal Opportunities</td>
<td>U</td>
<td>0</td>
<td>112.5</td>
<td>300</td>
<td>90.23</td>
</tr>
<tr>
<td>D3 Sustainable</td>
<td>U</td>
<td>0</td>
<td>156.25</td>
<td>300</td>
<td>97.97</td>
</tr>
</tbody>
</table>
4.4 Focus Group

RQ2b) Participants’ Alignment

Following the prioritization exercise, we carried out another activity aimed at assessing the same four main research quality areas and their corresponding sub-criteria; however using another technique. We displayed on a whiteboard the four main areas and corresponding sub-criteria, and ask participants to pick those areas/sub-criterion that they saw as most important and least important. To differentiate between these two types of decisions, participants were given green and red post-its – green symbolising most important and red least important. Figure 4.4 displays the results for both groups – Group 1 (8 participants), and group 2 (4 participants) are shown on the left and right hand sides, respectively.

Note that there are more dots in each area than the number of participants; this occurred because some participants considered more than one sub-criterion as equally important. Also note that this is a different assessment in which there was no normalisation of results; therefore, the assessment shown in the whiteboard does not map entirely to the normalized values of importance provided in Figure 4.3. In other words, the whiteboard presents the view of the participants with regard to a particular area (e.g., credible) and does not account for the importance of each sub-criteria weighted according to its main area, as described in Section 4.3.1.

An excellent example of this difference is related to the sub-criteria B2.1 relevant research idea. It received a large number of green markers (see Figure 4.4), but it is not ranked as the most important sub-criteria of the contributory dimension (Figure 4.3). In particular, the wider distribution of its box and right whisker suggests misalignment on the opinion of participants with regarding B2.1.

In order to summarize the results of the whiteboard, we applied vote counting to the markers, dividing the total of reds and green markers by the total number of participants so to obtain a normalized view, as detailed in Table A1.3. The interpretation of the results of the whiteboard exercise were categorized according to four different patterns, as follows:

**ABG) Alignment between-groups:** represents whether similar views (either in agreement or conflicting) emerged in both groups.

**AWG) Alignment within-groups:** represents whether most of the group participants agree on the importance of a particular sub-criteria.

**MBG) Misalignment between-groups:** represents whether there are conflicting views of importance between participants from different groups.
Figure 4.4: Representation of a whiteboard describing the most and least important dimensions, according to the focus group sessions. Each green dot represents a sub-criterion assessed as highly important by a participant; conversely, a red dot represents a sub-criterion assessed as least important.
### 4.4 Focus Group

MWG) **Misalignment within-group**: represents whether conflicting views of importance exist within participants of a group.

With regard to the alignment between-groups (ABG), we identified two major cases. First, A1.1 internally valid is considered important by most of the participants in both groups (more than 90% of participants). Second, B2.1 relevant research also received a lot of green markers from both groups. This result suggests a preference for methodological strictness and practical relevance.

Table A1.3: Summary of the answers from both focus group sessions at the whiteboard. In the columns labelled Green and Red we account for the amount of markers in relation to the group size, i.e. number of participants. The rightmost column describe our interpretation of the findings, according to three categories.

<table>
<thead>
<tr>
<th>Sub-criteria</th>
<th>Focus Group 1</th>
<th>Focus Group 2</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Green</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>A1.1</td>
<td>100% (8/8)</td>
<td>-</td>
<td>75% (3/4)</td>
</tr>
<tr>
<td>A1.2</td>
<td>87.5% (7/8)</td>
<td>-</td>
<td>25% (1/4)</td>
</tr>
<tr>
<td>A1.3</td>
<td>50% (4/8)</td>
<td>-</td>
<td>50% (2/4)</td>
</tr>
<tr>
<td>A2</td>
<td>12.5% (1/8)</td>
<td>-</td>
<td>50% (2/4)</td>
</tr>
<tr>
<td>A3</td>
<td>12.5% (1/8)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A4</td>
<td>37.5% (3/8)</td>
<td>12.5% (1/8)</td>
<td>25% (1/4)</td>
</tr>
<tr>
<td>B1.1</td>
<td>25% (2/8)</td>
<td>50% (4/8)</td>
<td>-</td>
</tr>
<tr>
<td>B1.2</td>
<td>25% (2/8)</td>
<td>62.5% (5/8)</td>
<td>-</td>
</tr>
<tr>
<td>B1.3</td>
<td>37.5% (3/8)</td>
<td>50% (4/8)</td>
<td>-</td>
</tr>
<tr>
<td>B2.1</td>
<td>75% (6/8)</td>
<td>-</td>
<td>75% (3/4)</td>
</tr>
<tr>
<td>B2.2</td>
<td>50% (4/8)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B2.3</td>
<td>50% (4/8)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B3</td>
<td>12.5% (1/8)</td>
<td>-</td>
<td>25% (1/4)</td>
</tr>
<tr>
<td>C1.1</td>
<td>62.5% (5/8)</td>
<td>25% (2/8)</td>
<td>25% (1/4)</td>
</tr>
<tr>
<td>C1.2</td>
<td>50% (4/8)</td>
<td>25% (2/8)</td>
<td>50% (2/4)</td>
</tr>
<tr>
<td>C1.3</td>
<td>50% (4/8)</td>
<td>25% (2/8)</td>
<td>100% (4/4)</td>
</tr>
<tr>
<td>C2</td>
<td>12.5% (1/8)</td>
<td>-</td>
<td>50% (2/4)</td>
</tr>
<tr>
<td>C3</td>
<td>12.5% (1/8)</td>
<td>-</td>
<td>25% (1/4)</td>
</tr>
<tr>
<td>D1</td>
<td>25% (2/8)</td>
<td>25% (2/8)</td>
<td>50% (2/4)</td>
</tr>
<tr>
<td>D2.1</td>
<td>75% (6/8)</td>
<td>12.5% (1/8)</td>
<td>50% (2/4)</td>
</tr>
<tr>
<td>D2.2</td>
<td>50% (4/8)</td>
<td>12.5% (1/8)</td>
<td>-</td>
</tr>
<tr>
<td>D2.3</td>
<td>37.5% (3/8)</td>
<td>37.5% (3/8)</td>
<td>-</td>
</tr>
<tr>
<td>D3</td>
<td>25% (2/8)</td>
<td>25% (2/8)</td>
<td>-</td>
</tr>
</tbody>
</table>
In some other cases, we noted aligned views of importance within a group (AWB), but that is not shared by the participants of the other groups. Most of the participants in group 1 considered the subcriteria A1.2 reliable as important. Participants from group 2 are more aligned towards C1.3 readable, the better ranked sub-criteria related to main area C communicable.

Misalignment between-groups are not easily explained, as the participants from different groups did not discuss the issue between them. We identified a relevant MBG case related to B2.3 current idea, within the contributory area.

Misalignments within-group (MWG) were more common, and they were also the ground for discussions during the focus group sessions. With regarding focus group 1, most of the MWG were identified concerning the original subcriteria (i.e., B1.1 to B1.3) and the main area conforming (i.e., D1 to D3). In relation to group 2, the MWG occurred about three subcriteria of the credible main area (i.e., A1.2, A1.3, and A2), two of the conforming dimension (C1.3 and C3) and another one concerning B3. Finally, two MWG were identified in both of the groups (C1.1 and D1), thus denoting alignment between-groups.

RQ2c) Summary of the participants’ opinions

Misalignments within-group were also ground for the discussions during the focus groups sessions. Whenever there was an MWG with regard to a main area’s or subcriterion’s importance (i.e., both green and red markers were present), participants were invited for discussion. The goal of such discussion was not to reach consensus but to understand in more detail their rationale. Notes were taken during such discussions; we also recorded the focus group sessions and used it to complement the notes. A summary of the gathered opinions is compiled in Table A1.4.

Besides the relevance of the dimensions, participants discussed how such model should be operationalized, and what should be the intent behind using such model to assess research quality. The group 1 discussions focused upon the context to which such quality dimensions apply. The opposing views of academia vs. industry, quantitative vs. qualitative research, research practice vs. research report, and the different philosophical stances (e.g., positivism, pragmatism) were mentioned.
Table A1.4: Summary of the participants’ opinions. The left column shows the dimensions of research quality and glossary items, and the right column presents short sentences extracted from the transcripts. For each sentence, we identify the participant or group that provided it, e.g. P1 refers to Participant 1 whereas G2 refers to the Focus Group 2.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Credible</td>
<td>It is more relevant to academia than industry (P7); it is also a pre-condition to assess other dimensions (P1)</td>
</tr>
<tr>
<td>A1 Rigorous</td>
<td>-</td>
</tr>
<tr>
<td>A1.1 Internally valid</td>
<td>The description is incorrect (P11) or narrow (P12)</td>
</tr>
<tr>
<td>A1.2 Reliable</td>
<td>The most important criteria (P5); its description can be improved (P9); it should be broken down into subcriteria (P5, P9, P11)</td>
</tr>
<tr>
<td>A1.3 Contextual</td>
<td>The description is vague, resulting in divergent opinions (G2); it depends of A1.1 Internally valid (P12)</td>
</tr>
<tr>
<td>A2 Consistent</td>
<td>-</td>
</tr>
<tr>
<td>A3 Coherent</td>
<td>Overlapping with A2 (P9)</td>
</tr>
<tr>
<td>A4 Transparent</td>
<td>Its importance depends of the publication venue (P1); the description is not clear (P11); it is also associated to replicability (P5, P9) and A1.2 Reliable (P12)</td>
</tr>
<tr>
<td>B Contributory</td>
<td>It is less important (G1) or equally important (G1) to A Credible</td>
</tr>
<tr>
<td>B1 Original</td>
<td>It is more relevant to academia than industry (P2, P7); should be prioritized in applied research (P8); low importance to replication studies (P5); its sub-criteria are not clear (G2)</td>
</tr>
<tr>
<td>B1.1 Original idea</td>
<td>-</td>
</tr>
<tr>
<td>B1.2 Original procedure</td>
<td>-</td>
</tr>
<tr>
<td>B1.3 Original results</td>
<td>'Provable knowledge' is not clear in the description (P9)</td>
</tr>
<tr>
<td>B2 Relevant</td>
<td>The most important criteria (P4); it should be prioritized over B1 Original (G1, G2)</td>
</tr>
<tr>
<td>B2.1 Relevant research idea</td>
<td>-</td>
</tr>
<tr>
<td>B2.2 Applicable result</td>
<td>-</td>
</tr>
<tr>
<td>B2.3 Current idea</td>
<td>The description is vague (P9); it should be removed (P12)</td>
</tr>
</tbody>
</table>

*Continued on next page*
### Table A1.4 – Continued from previous page

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3 Generalizable</td>
<td>It should also include the aspect of transferability (P9, P10, P12); it is related to both A Credible and B2 Relevant (P11)</td>
</tr>
<tr>
<td>C Communicable</td>
<td>It is an secondary dimension focused on the report only (P1, P8, P11); it depends of the target audience (P1, P4, P12); it is required to assess other dimensions (P12)</td>
</tr>
<tr>
<td>C1 Consumable</td>
<td>-</td>
</tr>
<tr>
<td>C1.1 Structured</td>
<td>C1.2 Understandable: It is redundant with C1.3 Readable (P3, P6, G2)</td>
</tr>
<tr>
<td>C1.3 Readable</td>
<td>-</td>
</tr>
<tr>
<td>C2 Accessible</td>
<td>It depends of the publisher (P12)</td>
</tr>
<tr>
<td>C3 Searchable</td>
<td>Overlapping with C2 Accessible (P3, P6)</td>
</tr>
<tr>
<td>D Conforming</td>
<td>It is an secondary dimension (P8); it has been given less priority (P6), but it is important for the evolution of SE research (P8)</td>
</tr>
<tr>
<td>D1 Aligned with Regulations</td>
<td>Represents a trade-off with innovative research (P4); it is impossible to conform to all regulations at once (P12)</td>
</tr>
<tr>
<td>D2 Ethical</td>
<td>It is a pre-condition to assess other dimensions (P1, P8); the description does not sort out between an ethical research process and ethics about the results and implications to practice (P2, P4)</td>
</tr>
<tr>
<td>D2.1 Morally justifiable</td>
<td>-</td>
</tr>
<tr>
<td>D2.2 Open</td>
<td>-</td>
</tr>
<tr>
<td>D2.3 Equal Opportunities</td>
<td>-</td>
</tr>
<tr>
<td>D3 Sustainable</td>
<td>‘System of rules’ is not clear in the description (P2, P7)</td>
</tr>
<tr>
<td>Glossary</td>
<td>It does not help to explain the quality dimensions (P11)</td>
</tr>
<tr>
<td>New knowledge</td>
<td>Does it means evidence? (P12)</td>
</tr>
<tr>
<td>Rules of description</td>
<td>Not clear what it means (P1, P2, P9, P11)</td>
</tr>
<tr>
<td>System of rules</td>
<td>Not clear what it means (P2, P7)</td>
</tr>
</tbody>
</table>

Both groups also pointed out descriptions of several dimensions that, in their view, were not clear. This was also true in relation to the glossary (see Appendix 4.8.2), which is provided in the original paper, and is supposed to facilitate understanding. Some of the participants suggested improvements to the phrasing and also to the model’s structure.
Both groups criticized the model’s structure. Further, one of group 1’s participants also mentioned that the process to cluster or aggregate the dimensions was questionable. Participants also pointed out redundancies (e.g., between C1.2 Understandable, and C1.3 Readable); dependencies (e.g., A1.3 Contextual depends on A1.1 Internally valid), and overlaps (e.g., B3 Generalizable seems to belong both to contributory and credible areas).

Finally, both groups also believed that the relationships between some dimensions were not explicit, and the dependencies should be made clearer. One participant suggested the use of pass-only standard, shaping the model into a series of decision points. We organized such suggestions and proposed a revised model of research quality (see Section 4.5.3).

4.5 Discussion

4.5.1 Reflections on Investigating Research Quality Alignment

Our research approach used the SAAMSQ, a method to determine the alignment of stakeholders proposed by Barney and Wohlin [463]. The method, originally designed for software quality standards, performed well under the circumstances of our study. In particular, it matches Mårtensson et al. [483] recommendations, in which the dimensions of quality should be discussed, weighted and prioritized by experts on the field the method was found to be suitable.

Hence, in further studies on quality alignment in the context of empirical software engineering we recommend to use the SAAMSQ. Similar experiences reported by Barney and Wohlin were made, namely that the degree of alignment becomes explicit and that conducting the study provides a foundation for understanding the underlying quality model. In our investigation, the prioritization made people reflect on the definitions in depth, which led to suggestions for changes to the quality model by Mårtensson et al. [483] to fit empirical software engineering.

Researchers willing to apply a similar approach to assess the alignment of different groups of experts are encouraged to follow our methodology. We briefly summarize the method and how to use it to further investigate quality alignment in empirical software engineering.

1. Select a case context. First, it is essential to ensure that the context in which the process will be applied is sufficiently narrow. In this article, we
Aligning the Views of Research Quality in Empirical Software Engineering

described two distinct applications, the first with researchers from SERL Sweden, and the second within the ISERN annual meeting (see Sections 4.3.1 and 4.4.1 for details). In both cases, we investigated the generic standards of research quality in ESE, but more specific topics (e.g., related to a particular philosophical stance or activity in the research process) are also interesting points for an investigation. As the investigation was conducted on two levels (research group and community) we also want to highlight that the model may be applied to a very specific research context and study.

2. **Identify key stakeholders.** To ensure that our candidates are interested in the investigated topic, we advertised the research idea and provided means for self-selection. The participants on both cases covered a wide range of research competencies and provided different perspectives regarding the topic.

3. **Select/tailor a quality model.** We selected the multidisciplinary framework for classification of the quality of research proposed by Mårtensson et al. [483] to represent the aspects of research quality. Alternative models, describing different quality concepts are available, e.g. [466, 477, 481]. However, given that the models are focused on specific quality concepts, we suggest to further use the updated and generic model presented in this paper (see Section 4.5.3) as a basis for future studies.

4. **Develop a data collection instrument.** We developed the data collection instrument as an excel spreadsheet grounded on previous works [480, 487]. The subjects assigned importance values using hierarchical cumulative voting (HCV), and the assigned value for each subcriterion was weighted according to the values assigned to its related main area.

5. **Conduct data collection.** We conducted workshop sessions to collect the selected participants’ opinion. In the case study, the workshops took place in a meeting room at the work environment and lasted from 40 to 60 minutes in length. In the focus group, we conducted two sessions in the ISERN annual meeting, lasting 90 minutes each. The focus group sessions also included a whiteboard representation of the participants’ reasoning and round-robin discussions.

6. **Analyze the results.** A detailed explanation of the data analysis procedure is given in [463]. We aggregated the relative values of importance into a shared data frame and used descriptive statistics to present the
4.5 Discussion

We interpreted the data according to predefined categories (see Section 4.3.1) and identified relevant patterns regarding alignment and misalignment.

7. Present the results to participants. The results should be presented to participants asking them for responses to follow-up questions. After the case study’s workshops, we showed the preliminary results for the participant’s audience and collected their feedback to improve our study. The feedback was incorporated into the focus group design, in which a revised model of research quality for software engineering was produced (see Section 4.5.3). Finally, we sent the resulting model to the participants and asked them for feedback.

4.5.2 Alignment with Regard to the Importance Values

In this article, we presented two distinct studies (i.e. a case study and a focus group), each of which aiming to collect the views of importance from different stakeholder groups. This resulted in two distributions of relative importance (see Figures 4.2 and 4.3). Although similarities can be noted, they represent the particular views of two different groups.

During the case study, we determined the alignment according to the distributions of assigned importance values, i.e. the length of boxplots and whiskers, and the relative position of its central point with regarding a threshold. We identified interesting patterns, but could not explain the reason for them. Based on the feedback provided by participants, we designed a more comprehensive study.

As a next step, we employed a focus group approach to further investigate the perspective of experts regarding the conceptual dimensions of quality. The importance values were summarized by the participants themselves, providing a visual description of potential alignments and misalignments. Further discussions elicited different views regarding the importance of quality dimensions, the conceptual model and its applicability.

The mixed media we adopted in the focus groups provided richer evidence in comparison to the prioritization questionnaire alone. Additional qualitative data (i.e. group session transcripts and notes) provided a deeper understanding of the alignment with regard to the importance of quality dimensions.

Participants also pointed out dependencies among the dimensions. We would like to conduct further studies to explore the interdependence between quality dimensions and identify potential trade-offs under different perspectives. We
also encourage researchers willing to apply and evolve our approach to report the results and experiences from future applications.

When comparing the results of both studies (case study and focus group), representing the research group level and community level respectively, the following observations were identified. Both the case study as well as the focus group resulted in the same order of priorities with regard to importance: 1. credible (A), 2. contributory (B), 3. communicable (C), and 4. conforming (D). Also, the variance (i.e. misalignment) on the four dimensions was lower than on the detailed level.

An interesting observation related to the degree of variance (i.e. misalignment) between research group perspective and community perspective. The spread of data on the community level (average $\sigma = 144.07$) is larger compared to the spread on the research group level (average $\sigma = 124.51$). As a research group, regular interactions with respect to discussing research and collaborations on research papers occur, which may lead to a more aligned view of research quality. This may show an indication for the need for an intensified dialog about the priorities and definitions of research quality on the community level. Such a dialogue may have important consequences. For example, if internal validity is considered the highest priority in the community overall, this may lead to a strategic decision to design and conduct studies that allow achieving a high degree of that validity (such as experimental studies in controlled environments).

### 4.5.3 Revised Model of Research Quality

During the focus group session, we explicitly asked the participants if they think the model proposed by Mårtensson et al. [483] express the dimensions of research quality. Most of them agreed, but they mentioned a few concerns regarding:

1. The structure of the model;
2. The description of the dimensions; and
3. Applicability of the model for research appraisal.

We further discussed these issues and elicited suggestions from the participants on how to address such challenges. To solve issues 1 and 2, we revised the model based on the suggestions and propose new descriptions grounded in the literature.
Our update mainly reflects the suggestions of the focus group session. We summarized the comments and addressed one dimension at a time, assessing whether the suggestions are essential. Dimensions that are classified as highly important or unimportant (Section 4.4.2) and the ones presenting misaligned opinions (Section 4.4.2) were given extra attention.

In some cases, suggestions from different participants conflicted, so we applied a decision-making process based on vote counting. We used the transcripts to assess how many participants mentioned and agreed with a particular suggestion. We kept the four main areas, as none participant understood that this is an essential issue to the model. Figure 4.5 shows the revised structure of the model, which include the following updates:

- **A1.2 Reliable**: broken down in two: A1.2a Methodological rigorous and A1.2b Reproducible. Methodological rigorous is described as “The research method is correctly used for its intended purpose” [466, 470, 473]. Reproducible is “The findings are likely the same if the research is repeated, e.g. by independent researchers, at a different time, in a different place” [492].

- **A5 Coherent**: merged into A4 Consistent; the new merged dimension is defined as “Consistent: The research problem, research questions and research methodology are linked” [470].

- **B1.1 to B1.3**: merged into B1 Original. The new Original dimension is described “The research addresses a new problem by means of original theory, methods and context” [468].

- **C1.2 Understandable**: merged into C1.3 Readable, which is now defined “The research documentation is able to be understood (in terms of tone, style, structure and semantics) by SE professionals” [466]; and

- **C5 Searchable**: merged into C4 Accessible; which is now “The research report is available and is easily found by the potential audience, in particular outside of academia” [481].

We also incorporated new descriptions for the dimensions that were mentioned vague or hard to understand. Fourteen out of the 26 items figuring the revised model have been rephrased as requested by the focus groups’ participants, as shown in Figure 4.5. In addition to their suggestions, we look up to methodological papers describing such quality dimensions to compose a new description.
Figure 4.5: Revised model of research quality grounded on the results of the focus group. The updates comprise a new hierarchical structure and novel descriptions of fourteen dimensions, listed on the right. Items marked with an asterisk (*) were rephrased according to suggestions of the participants.

We systematically identified a set of nine papers supporting a new description of the requested items, six of which are cited by Mårtensson et al. [483]. The papers cover a series of domains, such as Software Engineering [466,473,481,482], Research Methodology [470,484], Information Systems [468], Social Sciences [492] and Spatial Planning [467]. Besides that, we adopted two suggestions of the participants.
Overall, only few changes were suggested and hence it appears that the interdisciplinary model is suitable to be used in the context of empirical software engineering.

4.5.4 Using the Model to Appraise Research Quality

The original conceptual model [483] is intended to foster the development of research practice in multiple academic fields, by promoting a shared understanding of the quality dimensions. Suggested applications include appraising research reports, assessing research grants applications and support the creation of standards for science and innovation.

Before operationalization, the authors suggest that the model should be validated in the domain it is intended to be applied [483]. By conducting such validation, this study resulted in a revised model of research quality. To demonstrate how the model can be applied to research practice, we design an exemplary scenario:

Assuming we are willing to employ a checklist instrument to appraise the quality of research, we would like to investigate the comprehensiveness of the checklist with regard to the quality dimensions.

We opted for using an unified checklist [495] that combine a series of recommendations for experiments and case study research. We employed a coding process using the dimensions in the revised model of research quality (see Figure 4.5) as a codebook. Table A1.4 summarizes the checklist items and the dimensions we assigned to each of them. A more extensive description of the checklist items is provided by Wieringa [495].

Table A1.5: Unified Checklist for Empirical Research in Software Engineering by Wieringa [495]. The rightmost column shows the categorization of the checklist questions according to the revised model of research quality.

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Research problem investigation</strong></td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>What is the higher-level engineering cycle?</td>
<td>A1.3, B1</td>
</tr>
<tr>
<td>U2</td>
<td>Knowledge goal in that cycle?</td>
<td>A1.3, B1</td>
</tr>
<tr>
<td>U3</td>
<td>Conceptual model of the phenomena?</td>
<td>A1.3, A2</td>
</tr>
<tr>
<td>U4</td>
<td>Conceptual model validity? (including construct validity)</td>
<td>A1.1, A2</td>
</tr>
<tr>
<td>U5</td>
<td>Unit of study (population)?</td>
<td>A1.1</td>
</tr>
<tr>
<td>U6</td>
<td>Research questions?</td>
<td>B2.1</td>
</tr>
<tr>
<td>U7</td>
<td>Current knowledge?</td>
<td>B2.3</td>
</tr>
</tbody>
</table>

Continued on next page
Table A1.5 – Continued from previous page

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Research design</strong></td>
<td></td>
</tr>
<tr>
<td>U8</td>
<td>Unit of data collection? (sample, model or case)</td>
<td>A1.1, A1.2a</td>
</tr>
<tr>
<td></td>
<td>U8.1 Acquisition?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U8.2 Structure?</td>
<td></td>
</tr>
<tr>
<td>U9</td>
<td>Treatment of unit of data collection?</td>
<td>A1.1, A1.2a</td>
</tr>
<tr>
<td></td>
<td>U9.1 Treatment specification?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U9.2 Treatment assignment?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U9.3 Treatment plan?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U9.4 Treatment instruments?</td>
<td></td>
</tr>
<tr>
<td>U10</td>
<td>Measurement of unit of data collection?</td>
<td>A1.1, A1.2a</td>
</tr>
<tr>
<td></td>
<td>U10.1 Measurement procedures?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U10.2 Measurement instruments?</td>
<td></td>
</tr>
<tr>
<td>U11</td>
<td>Kind of reasoning? (statistical or case-based)</td>
<td>A1.2a</td>
</tr>
<tr>
<td></td>
<td><strong>Research design validation</strong></td>
<td></td>
</tr>
<tr>
<td>U12</td>
<td>Validity of unit of data collection?</td>
<td>A1.1</td>
</tr>
<tr>
<td></td>
<td>U12.1 External validity?</td>
<td>B3</td>
</tr>
<tr>
<td></td>
<td>U12.2 Ethics?</td>
<td>D1</td>
</tr>
<tr>
<td>U13</td>
<td>Validity of treatment?</td>
<td>A1.1</td>
</tr>
<tr>
<td></td>
<td>U13.1 Instrument validity?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U13.2 External validity?</td>
<td>B3</td>
</tr>
<tr>
<td></td>
<td>U13.3 Ethics?</td>
<td>D1</td>
</tr>
<tr>
<td>U14</td>
<td>Validity of measurement?</td>
<td>A1.1</td>
</tr>
<tr>
<td></td>
<td>U14.1 Validity of measurement procedures?</td>
<td>A1.2a</td>
</tr>
<tr>
<td></td>
<td>U14.2 Instrument validity?</td>
<td>A1.1</td>
</tr>
<tr>
<td>U15</td>
<td>Validity of reasoning?</td>
<td>A1.2a</td>
</tr>
<tr>
<td></td>
<td>U15.1 Conclusion validity?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U15.2 Internal validity?</td>
<td>A1.1</td>
</tr>
<tr>
<td></td>
<td><strong>Research execution</strong></td>
<td></td>
</tr>
<tr>
<td>U16</td>
<td>Unit of data collection?</td>
<td>A4</td>
</tr>
<tr>
<td></td>
<td>U16.1 Acquisition?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U16.2 Quality?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U16.3 History?</td>
<td></td>
</tr>
<tr>
<td>U17</td>
<td>Execution of treatment?</td>
<td>A1.2b, A4</td>
</tr>
<tr>
<td>U18</td>
<td>Execution of measurements?</td>
<td>A1.2b, A4</td>
</tr>
<tr>
<td>U19</td>
<td>Availability of data?</td>
<td>A4</td>
</tr>
<tr>
<td>U20</td>
<td>Provenance of data?</td>
<td>A4</td>
</tr>
</tbody>
</table>

*Continued on next page*
The checklist is structured in five sections mapping the phases of the empirical cycle [495]. The first section focus on the conceptualization of the study and the definition of the research problem to be investigated. Quality aspects herein are focused on rigorous method (A1), original and relevant contributions (B1 and B2).

The research design section concerns the proper use of research methods (A1.2a) to correctly measure the phenomenon (A1.1) it is sought to investigate. The next section addresses the validation of such design, and thus cover similar aspects, also, to ensure that the results are generalizable or transferable to other contexts (B3).

Research execution is focused on trustfully documenting the research process (A1.2a), including any unexpected event that could influence the results. Transparency (A4) is required for further assessment and replication of the study. Finally, the results evaluation section focus on communicating the research (C1) aiming to provide a meaningful contribution (A2.2) to the field.

The categorization of the checklist includes all four main areas of research quality, with different degrees of coverage. Areas (A) credible and (B) contributory are more expressed than (C) communicable and particularly, (D) conforming. In particular, questions addressing the rigorous aspect (A1) are plenty, characterizing the checklist towards the methodological quality of research.

About the communicable dimension, the checklist does not address the readability and accessibility of the report. With regard to the conforming aspect, only D2 ethics is explicitly mentioned, with no specification of its subcriteria. D1 alignment to regulations and D3 sustainability are not covered.

The categorization of the checklist items is consistent with the results of importance assessment in both our case study and focus group (Sections 4.3.2
and 4.4.2). In this sense, one can assume a match between the proposed checklist instrument [495] and the views of importance regarding the research quality.

We also identified gaps related to the communicable (C) and conforming (D) dimensions of research quality. On the one hand, these two aspects are perceived as secondary, as they less concerned the research practice. On the other hand, their importance is grounded on the means researchers use to reach the potential audience and an ongoing debate on the ethical/moral orientation of research.

4.6 Conclusion

This paper reports a mixed method empirical study to validate a multidisciplinary framework of research quality. The study is aimed at capturing the views of researchers of two groups (SERL-Sweden and ISERN community) with regard to the importance of evaluating research quality.

We conducted a preliminary case study to determine quality alignment among stakeholders within a particular research group. During a series of workshops, we discussed a conceptual model of research quality and employed an electronic form to collect relative values for importance for quality dimensions. We validated the data collection instrument and obtained insights for complementary research.

We further conducted a focus group with experts from the ISERN community. The focus group employed a design alike to our case study research, but focused on the qualitative opinions of participants regarding the nature of research quality. The participants acknowledged the importance of the research quality dimensions. There were, however, concerns regarding (1) the structure of the model, (2) the description of the dimensions, and (3) applicability of the model to research appraisal.

Based on the results from our mixed approach, the following conclusions were drawn:

RQ1a/RQ2a What are the most important research quality dimensions in SE research, as per the participants?

In both the case study and focus group, the main areas of research quality credible (A) and contributory (B) are ranked more important than communicable (C) and conforming (D). This reflects the orders of priorities for some of the subcriteria associated to credibility: A1.1 internally valid is ranked highly important by both groups, whereas A1.2 reliable, A1.3
4.6 Conclusion

contextual and A2 consistent are ranked at least important. With regard to contributory (B), B2.1 relevant research idea and B2.2. applicable result are considered important. Despite that, priorities for other subcriteria differ when comparing the results from both studies, e.g. B3 generalizable (B3) is ranked unimportant in the case study and important in the focus group.

RQ1b/RQ2b What is the level of alignment between the different participants with regard to their assessment of the importance of quality dimensions for SE research? Views of importance are diverse, and therefore misalignment among the participants are not unusual. In particular, the variance in the importance values points out misalignments with regard to the contributory (B) subcriteria B2.1 relevant research idea, B2.2. applicable results and B3. generalizable. Interestingly, the alignment is better with regard to the subcriteria ranked less important. Moreover, the alignment is higher among participants from the case study (a research group perspective) than from the focus group (community perspective).

In the focus group, we observed that misalignments are more frequent than alignments, both within the participants of the same group session and across different group sessions. We also discussed the misalignments with participants to identify reasons and needs for improvement.

RQ2c To what extent the model of research quality is relevant for software engineering research? According to the focus groups, the interdisciplinary model proposed by Mårtensson et al. [483] fairly express the dimensions of research quality, and it has a potential to be operationalized in the context of ESE research. Limitations regarding the structure of the model and the description of the criteria were addressed following the opinions of experts. As a result, we produced a revised model of research quality and provided an example of use to characterize a research assessment instrument.

As future work, we aim to replicate the process to determine the alignment of stakeholders involved in different aspects of research practice, e.g. reviewers, research institutions, funding agencies. By doing so, we expect to evaluate and evolve our revised model taking into consideration a more comprehensive view of ESE research. Furthermore, inspired by the exemplary scenario presented in Section 4.5.4, we intend to operationalize the revised model of research quality to create and to validate instruments to assess the quality of research.
4.7 References


Aligning the Views of Research Quality in Empirical Software Engineering


## 4.8 Appendix

### 4.8.1 Definitions of the quality dimensions according to Mårtensson et al. [483]

<table>
<thead>
<tr>
<th>#</th>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Credible</td>
<td>Research that is Coherent, Consistent, Rigorous and Transparent</td>
</tr>
<tr>
<td>A1</td>
<td>Rigorous</td>
<td>Research that is Contextual, Internally Valid and Reliable</td>
</tr>
<tr>
<td>A1.1</td>
<td>Internally valid</td>
<td>A correct Scientific Method (incl. research design) is used in relation to the Question at Hand and Context, and New Knowledge is Provable.</td>
</tr>
<tr>
<td>A1.2</td>
<td>Reliable</td>
<td>The chosen Scientific Method is appropriate for the present Question at Hand and Context, and is documented in a Described Procedure that others could use to reach a similar result in the same Context.</td>
</tr>
<tr>
<td>A1.3</td>
<td>Contextual</td>
<td>Existing Knowledge that is relevant for the Context is used, and is presented according to Rules for Description</td>
</tr>
<tr>
<td>A2</td>
<td>Consistent</td>
<td>New Knowledge is logically linked to Existing Knowledge and is in accordance with the Scientific Method and Question at Hand</td>
</tr>
<tr>
<td>A3</td>
<td>Coherent</td>
<td>Adequate consideration is given to Existing Knowledge in the chosen Context.</td>
</tr>
<tr>
<td>A4</td>
<td>Transparent</td>
<td>Relevant New Knowledge in the reporting of research results is included and the process is described in relation to the Question at Hand, Scientific Method and Existing Knowledge</td>
</tr>
<tr>
<td>B</td>
<td>Contributory</td>
<td>Research that is Original, Relevant and Generalizable</td>
</tr>
<tr>
<td>B1</td>
<td>Original</td>
<td>Research that has an Original Idea, uses an Original Procedure and produces an Original Result</td>
</tr>
<tr>
<td>B1.1</td>
<td>Original idea</td>
<td>The Question at Hand has not been asked before in the current Context or is interpreted in a novel way</td>
</tr>
<tr>
<td>B1.2</td>
<td>Original procedure</td>
<td>Described Procedure is original in relation to the Question at Hand</td>
</tr>
<tr>
<td>B1.3</td>
<td>Original result</td>
<td>New Knowledge is Provable in relation to Existing Knowledge</td>
</tr>
<tr>
<td>B2</td>
<td>Relevant</td>
<td>Research that has a Relevant Research Idea, Applicable Result and Current Idea</td>
</tr>
</tbody>
</table>
# Table A1.6 – Continued from previous page

<table>
<thead>
<tr>
<th>#</th>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2.1</td>
<td>Relevant research idea</td>
<td>Question at Hand is relevant for the current Target Group</td>
</tr>
<tr>
<td>B2.2</td>
<td>Applicable result</td>
<td>New knowledge is Beneficial for the current Target Group</td>
</tr>
<tr>
<td>B2.3</td>
<td>Current idea</td>
<td>The Question at Hand is in accordance with the current Context</td>
</tr>
<tr>
<td>B3</td>
<td>Generalizable</td>
<td>New Knowledge is practically or theoretically useful in Contexts other than the one studied</td>
</tr>
<tr>
<td>C</td>
<td>Communicable</td>
<td>The research is consumable, accessible and searchable</td>
</tr>
<tr>
<td>C1</td>
<td>Consumable</td>
<td>Research that is Structured, Understandable and Readable</td>
</tr>
<tr>
<td>C1.1</td>
<td>Structured</td>
<td>The Research documentation follows the Rules for Description</td>
</tr>
<tr>
<td>C1.2</td>
<td>Understandable</td>
<td>The language in the Research documentation is understandable for the Target Group</td>
</tr>
<tr>
<td>C1.3</td>
<td>Readable</td>
<td>Correct language in the Research documentation for the Target Group</td>
</tr>
<tr>
<td>C2</td>
<td>Accessible</td>
<td>New Knowledge is easily available to the Target Group</td>
</tr>
<tr>
<td>C3</td>
<td>Searchable</td>
<td>The documented New Knowledge is structured according to the Rules for Description and easily found by the Target Group</td>
</tr>
<tr>
<td>D</td>
<td>Conforming</td>
<td>The research is aligned with regulations, ethical and sustainable.</td>
</tr>
<tr>
<td>D1</td>
<td>Ethical</td>
<td>The Research is Morally Justifiable, Open and supports Equal Opportunities</td>
</tr>
<tr>
<td>D1.1</td>
<td>Morally justifiable</td>
<td>The Research complies with currently applicable ethical standards as described in the System of Rules</td>
</tr>
<tr>
<td>D1.2</td>
<td>Open</td>
<td>The Research demonstrates Transparency with currently applicable ethical standards as described in the System of Rules</td>
</tr>
<tr>
<td>D1.3</td>
<td>Equal opportunities</td>
<td>The Research is consistent with equal treatment according to the System of Rules</td>
</tr>
<tr>
<td>D2</td>
<td>Aligned with regulations</td>
<td>The Research complies with currently applicable legal aspects of the System of Rules</td>
</tr>
<tr>
<td>D3</td>
<td>Sustainable</td>
<td>The Research complies with sustainable development aspects as described in the System of Rules</td>
</tr>
</tbody>
</table>
4.8.2 Definitions of concepts in the model of research quality according to Mårtensson et al. [483]

<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor</td>
<td>A Person initiating and/or performing a Conscious Action</td>
</tr>
<tr>
<td>Beneficial</td>
<td>A positive effect of New Knowledge for a Target Group</td>
</tr>
<tr>
<td>Conscious action</td>
<td>A process initiated and/or performed by an Actor</td>
</tr>
<tr>
<td>Context</td>
<td>An environmental or intellectual setting where the Research takes place and/or is studied, and where Existing Knowledge is valid</td>
</tr>
<tr>
<td>Described procedure</td>
<td>A description of how the Research will be performed and documented according to the Rules for Description</td>
</tr>
<tr>
<td>Existing knowledge</td>
<td>Knowledge that is built on by the Research, exists in a Context, can be documented in a Source and is expanded with New Knowledge</td>
</tr>
<tr>
<td>New knowledge</td>
<td>Knowledge that expands Existing Knowledge, is Provable, and is Beneficial for a Target Group</td>
</tr>
<tr>
<td>Person</td>
<td>A human being</td>
</tr>
<tr>
<td>Provable</td>
<td>Evidence that the New Knowledge is demonstrable</td>
</tr>
<tr>
<td>Question at hand</td>
<td>A research question that is the base for Research</td>
</tr>
<tr>
<td>Relationship</td>
<td>A relation between two Conscious Actions showing how those actions interact</td>
</tr>
<tr>
<td>Research</td>
<td>A Conscious Action that aims for New Knowledge, emanates from one or several Questions at Hand, studies one or several Contexts, builds upon Existing Knowledge, uses one or several Scientific Methods, is documented in one Described Procedure, requires Transparency and relates to one or several Systems of Rules</td>
</tr>
<tr>
<td>Rules for description</td>
<td>Rules describing what a Described Procedure should contain, including its intentions and results. This can differ in regards to Context, Scientific Method, System of Rules, Existing Knowledge and Question at Hand</td>
</tr>
<tr>
<td>Scientific method</td>
<td>A described and precise technique used for conducting the Research</td>
</tr>
<tr>
<td>Source</td>
<td>Documents, databases or other media that contain Existing Knowledge</td>
</tr>
<tr>
<td>System of rules</td>
<td>Legal requirements, regulations, norms and other guidelines that influence how Research should be performed</td>
</tr>
<tr>
<td>Target group</td>
<td>Individuals, organizations, enterprises and/or society that benefit from New Knowledge</td>
</tr>
<tr>
<td>Transparency</td>
<td>A clear description required by the Research</td>
</tr>
</tbody>
</table>
Chapter 5

Teaching students critical appraisal of scientific literature using checklists

This chapter is based on the following paper:


5.1 Introduction

In the age of rapid technological advancement, to stay abreast, students are required to have the ability to analyze problems, and to conceptualize, implement and evaluate new solutions [529]. This requires a scientific attitude and a disposition towards research in Software Engineering (SE) graduates.

The scientific attitude means having a systematic, skeptical and ethical approach [541]. The role of developing these skills of reflection is well recognized for software engineering research and practice [507, 515]. One way of attaining this is by incorporating education on research in the curriculum [516].
Some existing work in SE education focuses more broadly on the approach to combine research tasks with other teaching formats like lectures and seminars [527, 528]. However, in this study, we evaluate whether students with a little background and formal training in empirical research can critically appraise articles reporting experimentation and case study research.

To support students for this task of critical appraisal, we have made use of checklists. Originally, these checklists were intended to assess the quality of empirical research (e.g., [505, 520, 542, 550, 553]).

Explicit criteria help in objectifying and thus achieving a high agreement among reviewers [537] e.g., when selecting papers for inclusion in conference. Our proposition is that the novice researchers could benefit from the application of checklists. In particular, checklists support a systematic and more objective critical appraisal [549]. One may also argue that novices may depend more on the support by checklists and guidelines as they cannot rely on the same degree of review experience as their experienced peers [543].

In this paper, we explored whether two checklists proposed in the software engineering literature (one for experiments [553] and one for case study research [542]) aid novices (M.Sc. students in software engineering/computer science) in appraising the quality of studies. To achieve this aim, we made the following contributions:

C1) We assessed the accuracy of the students’ assessments with regard to a gold standard, to reflect the usefulness of the checklists to support novices to appraise the quality of empirical studies.

C2) We compared the concordance between the students’ assessment, to investigate whether the checklists are consistently interpreted and applied by students.

C3) We collected the opinions of the students about the checklists, in particular on the understandability and ease of use of the checklists.

C4) We have also used and assessed the pedagogical value of using the checklists as an instrument to teach students critical appraisal of scientific literature.

This paper is organized as follows: Section 6.2 summarizes related research; Section 5.3 describes the experimental design of our evaluation study; Section 6.3.3 presents the data and findings for each research question; we further discuss the findings in Section 5.5. Finally, Section 6.6 describes our conclusions and directions for future work.
5.2 Background

In this section, we present the main concepts related to our work: i) critical appraisal of scientific literature, ii) instruments supporting this activity in SE, and iii) the importance of attaining such knowledge.

5.2.1 Appraisal of Study Quality

The peer-review process is an essential part of the research practice. It comprises the assessment or evaluation of research work by qualified peers i.e., fellow researchers [544]. The process is often criticized for the difficulty to produce objective and reliable reviews [530,535,545].

Often more than one reviewer is employed to assess a particular research work, which requires a degree of agreement with respect to each other. Different understandings of the same work are likely to produce inconsistent reviews [530, 545]. Moreover, novice researchers often lack the knowledge and training needed to assess others’ work comprehensively [537,543]. In an attempt to minimize such issues, guidelines and checklists have been proposed to support reviewers [513,514,532,543,546].

Checklists are instruments that support task management by providing scales that, if properly followed, ensure that the task is conducted consistently and completely. They are applied in a wide range of domains, such as construction, safety procedures, medical care, and personality evaluation [499]. They can also support the appraisal of quality on diverse objects, such as research work.

5.2.2 Assessment Checklists in Software Engineering

In the context of this study, we are particularly interested in the checklists appraising the quality of research in SE. To assess the produced evidence, SE researchers employ checklists for reviewing empirical studies. These instruments are often related to a specific research method, such as case study [522,539,542], experiment [505,520,553] and systematic literature review [506,524]. Moreover, checklists addressing multiple methods have also been proposed [503,550,551].

It is important to ensure that the checklists are appropriate to assess the empirical studies. Kitchenham et al. [523] demonstrated that the reliability of such instruments depends on how they are applied, e.g. the assessment produced by two reviewers employing a discussion round is significantly more reliable than the one produced by an individual. Condori-Fernandez et al. [503] evaluation of a unified checklist identified issues related to its understandability
and applicability, e.g. reviewers’ ability to answer decrease when the checklist is too long.

5.2.3 Importance of Critical Appraisal in SE education

The ability of critically analyze scientific literature is required to researchers and potential reviewers, particularly on evidence-based practice. Ideally, such skill can be attained through critical appraisal activities [549].

For this purpose, Evidence-based Software Engineering (EBSE) approach has been introduced in the educational context [518,540]. EBSE attempts to influence SE practice such that the decisions regarding technology adoption and process improvement are based on evidence [525]. As the students from this course, will likely pursue careers as practitioners and researchers in the field of software engineering, it is imperative that they can assess the strength of evidence different studies present.

5.3 Research method

In the Master’s course on research methods, from the pedagogical perspective, we carry out critical appraisal activities to help students to develop analytical skills regarding empirical research. For this purpose, we relied on the use of checklists to support students to systematically review scientific literature. To assess how well checklists support students on the critical appraisal, we answered the following questions:

RQ1) How accurate are the students’ assessment when supported with checklists? We investigate this by comparing the students’ ratings to the ones produced by more experienced researchers. This is particularly related to the contribution C1.

RQ2) How consistent are the students’ assessment when supported with checklists? We visualized the students’ rates comparatively to identify concordances. This question relates to contribution C2.

RQ3) What are the students’ perception regarding the use of the checklists? We also collected qualitative data regarding the perception of the checklists (see the form in Appendix 5.8.3). The data were further compared to the students’ assessment to provide insights for RQ1 and RQ2. This is further associated with contributions C3 and C4.
5.3 Research method

5.3.1 Context

The study was conducted in the context of a post-graduate research methods course. The students in the course are from the M.Sc. in software engineering and computer science (CS) programs, both programs comprise 120 ECTS\textsuperscript{1} each. In both programs, a semester comprises 30 ECTS. The structure for the software engineering program is shown in Figure 5.1. As is visible, the research methodology course takes place in the second half of the second semester in both programs. That is, the students already participated in six courses within the program.

The course was intended to enable the students to describe the lifecycle of a research project from inception to completion. Beyond providing skills necessary for a researcher, this course attempts to inculcate an objective approach that fosters reflective practice and enable lifelong learning for graduates who will take roles in industry. Within this course, students are familiarized with academic literature, which is needed to solve the assignments. As the program is on the M.Sc. level, students are also required to write reports with discussions and their own reflections, and to synthesize knowledge [501].

The intended outcomes of the course after successful completion are, the students would be able to identify, evaluate and bring into service scientific literature and use an objective approach to solve practical problems in the potential roles of both academics and researchers.

The master’s programs are research focused with a strong focus on empirical components. The research methods course is seen as a preparation for the master’s thesis, which demands extensive research. Many students gained sufficient knowledge to, with the support of their supervisors, publish their M.Sc. thesis work in reputable SE journals (see e.g., [508,536,547]). Research methodology course is also formally a prerequisite for the M.Sc. thesis in Semester 4 (as shown in Figure 5.1)

5.3.2 Participants of the study

One of the important objectives of the course is developing the ability to critically appraise and discuss validity threats and limitations of a research study. For this purpose, in one of the assignments in the course, students are provided a set of papers to assess using the checklists. The data used in this study was generated in this mandatory assignment in the course.

\textsuperscript{1}European Credit Transfer System
Teaching students critical appraisal of scientific literature using checklists

<table>
<thead>
<tr>
<th>Semester 1</th>
<th>Semester 2</th>
<th>Semester 3</th>
<th>Semester 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seminar Series in Software Engineering (7.5 ECTS)</td>
<td>Agile and Lean Software Development (7.5 ECTS)</td>
<td>Software Architecture and Quality (7.5 ECTS)</td>
<td>Requirements Eng. and Product Mgt. (7.5 ECTS)</td>
</tr>
<tr>
<td>Software Architecture and Quality (7.5 ECTS)</td>
<td>Evidence-based Process change and evaluation (7.5 ECTS)</td>
<td>Requirements Eng. and Product Mgt. (7.5 ECTS)</td>
<td>Elective SE/CS (7.5 ECTS)</td>
</tr>
<tr>
<td>Research Methodology SE/CS (7.5 ECTS)</td>
<td>Elective SE/CS (7.5 ECTS)</td>
<td>Elective CS/SE (7.5 ECTS)</td>
<td>Software Evolution and Maintenance (7.5 ECTS)</td>
</tr>
<tr>
<td>Elective CS/SE (7.5 ECTS)</td>
<td>Software Evolution and Maintenance (7.5 ECTS)</td>
<td>Elective CS/SE (7.5 ECTS)</td>
<td>Software quality management (7.5 ECTS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M.Sc. Thesis (30 ECTS)</td>
</tr>
</tbody>
</table>

Figure 5.1: Overview of the SE program

38 groups (each consisting of two students) participated in the assignment. All the participants were master students from computer science and software engineering domain. Students were asked to select their peer and report their group through the learning management system used in the course.

Before the assignment, the students have had eleven (each 3 hours long) lectures on introduction to research, empirical research, research ethics, scientific
writing, experimentation, case study research, surveys, literature review, and simulation. Order of lectures are presented in Table A1.1. Along with these lectures, students were required to go through the guidelines for conducting and reporting the experiment and case study research (e.g., [517, 542, 553]). It was also required that students should read the experiment and case study research examples.

Table A1.1: Course overview and the context of the assignment

<table>
<thead>
<tr>
<th>Week</th>
<th>Lecture / Assignment</th>
</tr>
</thead>
</table>
| 3    | L1) Course Introduction  
      | L2) Introduction to Research Methods                                                 |
| 4    | L3) Overview of Research Methods  
      | L4) Ethics                                                                           |
| 5    | L5) Experiment  
      | L6) Problem Formulation  
      | L7) Case Study and Action Research  
      | L8) Survey-based Research                                                            |
| 6    | L9) Simulation  
      | L10) Searching for Scientific Literature  
      | L11) Literature Reviews  
      | A1) Critical Appraisal                                                                |
| 7    | L12) Data Analysis  
      | L13) Statistical Analysis + Introduction to R                                         |
| 10   | L14) Feedback related to (L6) Problem Formulation                                     |

5.3.3 Choice of objects (articles)

We decided to have a mix of articles with a known notion of quality. This was done to have a variety that mimics real-world artifacts. We chose three articles reporting experiments and three articles reporting case study research. We have only chosen journal articles because due to space limitation conference articles cannot report the methodology pursued in the study in sufficient detail.

The articles reporting experiments [498, 519, 531] were chosen from a set used by Kitchenham et al. [523]. They had assessed the papers using a checklist. We
used the reported ratings to label the papers accordingly regarding perceived quality as high, medium and low.

The three articles [504,509,534] reporting case study research were obtained through a systematic search and selection from the Scopus database of peer-reviewed literature. We collected a sample of five articles on topics covered by the disciplines of the SE program (see Figure 5.1) already completed by the students. Similarly to Kitchenham et al. [523], we selected three papers with a variety of perceived quality (i.e., high, medium and low).

### 5.3.4 Choice of interventions (checklists)

We identified several checklists for assessing empirical research in software engineering (e.g., [505,520,521,523,550]). For the study we decided to adopt the checklists prescribed in the two highly cited guidelines for conducting and reporting case study research [542] and experiments [553] in SE. Another reason for using these checklists was the students’ familiarity with the two methodological guidelines [542, 553], as these are part of the course literature. The checklists items are detailed in appendices 5.8.1 and 5.8.2.

We have classified the checklist items according to the main stages of experimentation and case study research [553]: P1) scope and context; P2) design and planning; P3) operation or data collection; P4) data analysis; and P5) presentation, packaging and reporting, as shown in Table A1.2. The classification allows us to analyze the data under different perspectives, e.g. a lack of consistency of the students’ ratings in a particular phase could indicate a pedagogical need.

Table A1.2: Classification of Checklist Items according to Empirical Research Phases

<table>
<thead>
<tr>
<th>Checklists</th>
<th>Case study</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1) scope and context</td>
<td>Q1 – Q3</td>
<td></td>
</tr>
<tr>
<td>P2) design and planning</td>
<td>Q1 - Q4</td>
<td>Q4 - Q12</td>
</tr>
<tr>
<td>P3) operation / data collection</td>
<td>Q5 - Q6</td>
<td>Q13 - Q16</td>
</tr>
<tr>
<td>P4) data analysis</td>
<td>Q7 - Q10</td>
<td>Q17 - Q19</td>
</tr>
<tr>
<td>P5) presentation and reporting</td>
<td>Q11 - Q12</td>
<td>Q20 - Q23</td>
</tr>
</tbody>
</table>

We believe that for most research phases (Table A1.2) the students can provide a fair assessment as they have been instructed on the theory of empirical
5.3 Research method

research. The scope and context phase, however, contains highly subjective questions that depend on the experience of a reviewer in the particular field of research, which the students have not yet obtained (see Q1, Q2, and Q3 in appendix 5.8.2). These questions are research-topic specific and students are likely not at the level of an expert reviewer. Therefore, we have focused on the analysis of questions about the rigor of methodology.

5.3.5 Assignment of object, subjects and treatments

Table A1.3 shows the possible combinations of objects that may be evaluated by each student group. For pedagogical reasons (as detailed in Section 5.5.3) we decided to disregard AX, BY and CZ. The remaining combination of a case study and an experiment were assigned to the 38 groups randomly. We ensured that each combination of experiment and case study was assigned to at least six groups. Due to the number of groups i.e., 38 CY and AZ were assigned to seven groups instead of six.

Table A1.3: Possible combination of treatments

<table>
<thead>
<tr>
<th>Case study Quality</th>
<th>Experiment Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>X</td>
</tr>
<tr>
<td>Medium</td>
<td>Y</td>
</tr>
<tr>
<td>Low</td>
<td>Z</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
</tr>
<tr>
<td>Medium</td>
</tr>
<tr>
<td>Low</td>
</tr>
</tbody>
</table>

5.3.6 Gold Standard

A gold standard is a measure of reference by which it is possible to assess the accuracy or conformance of others [548]. In the context of this study, the gold standard comprises the assessment produced by more experienced researchers that will be further compared to students'.

To produce such measure, two of the authors followed an approach similar to the students’ critical appraisal (Section 5.3.5) to assess all the articles (i.e., A, B, C, X, Y, and Z). The researchers discussed and reflected on each checklist
Teaching students critical appraisal of scientific literature using checklists

item in a way very alike to that expected from the students. The resulting scores are further used as a reference for comparison with the student’s average scores for each individual item.

Although the gold standard does not represent the truth due to the subjective nature of interpreting the checklist [525], we believe it fairly represents the judgment of experienced researchers (please see Section 5.5.3 for a more detailed discussion of validity implications).

5.3.7 Data Analysis

In this section, we briefly discuss the statistical methods that we used in our analysis. Firstly, we tried a robust statistical measure, Fleiss’ Kappa [502, 512] to compute accuracy and consistency values. However, since the Kappa value is highly influenced by the prevalence of the outcome, the two paradoxes, as highlighted by Feinstein & Cicchetti [511], appeared in our analysis. Thus, very low values of Kappa were recorded even when the absolute agreement among the students is high.

Therefore, we opted for reporting the descriptive statistics only. They involved the computation of simple summary statistics, like the percent agreement [533]. More specifically, we calculated the accuracy rate i.e., the percent agreement between the students’ rating and the gold standard, and the consistency i.e., the percent agreement across the students’ ratings. The statistics are accompanied by graphical representations like bar plots and horizontal bars divided by different colors.

5.4 Results

This section presents the analysis of students’ critical appraisal task according to the three research questions (see Section 5.3).

RQ1. Accuracy of Students’ Assessment

First, we compare students’ ratings to the gold standard to assess their accuracy. Tables A1.4 and A1.5 summarize the accuracy of students’ ratings of the checklists for case study and experiment, respectively.

The results are aggregated according to the objects (e.g., articles A, B, and C reporting case studies). For each checklist item, we present the gold standard rates (std.) and the accuracy (acc.) i.e., the percentage of the students’ ratings
### 5.4 Results

Table A1.4: Accuracy of ratings for Case Studies articles

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>0.67</td>
<td>Yes</td>
<td>0.75</td>
</tr>
<tr>
<td>Q2</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>0.83</td>
<td>Yes</td>
<td>0.75</td>
</tr>
<tr>
<td>Q3</td>
<td>Yes</td>
<td>0.8</td>
<td>N/A</td>
<td>0.5</td>
<td>N/A</td>
<td>0.5</td>
</tr>
<tr>
<td>Q4</td>
<td>Yes</td>
<td>0.8</td>
<td>Yes</td>
<td>0.83</td>
<td>Yes</td>
<td>0.92</td>
</tr>
<tr>
<td>Q5</td>
<td>Yes</td>
<td>0.9</td>
<td>Yes</td>
<td>0.83</td>
<td>Yes</td>
<td>0.5</td>
</tr>
<tr>
<td>Q6</td>
<td>Yes</td>
<td>0.8</td>
<td>Yes</td>
<td>0.75</td>
<td>No</td>
<td>0.42</td>
</tr>
<tr>
<td>Q7</td>
<td>Yes</td>
<td>0.6</td>
<td>Yes</td>
<td>1</td>
<td>No</td>
<td>0.67</td>
</tr>
<tr>
<td>Q8</td>
<td>Yes</td>
<td>0.8</td>
<td>Yes</td>
<td>1</td>
<td>No</td>
<td>0.67</td>
</tr>
<tr>
<td>Q9</td>
<td>Yes</td>
<td>0.7</td>
<td>No</td>
<td>0.67</td>
<td>No</td>
<td>0.33</td>
</tr>
<tr>
<td>Q10</td>
<td>Yes</td>
<td>0.8</td>
<td>No</td>
<td>0.58</td>
<td>Yes</td>
<td>0.67</td>
</tr>
<tr>
<td>Q11</td>
<td>Yes</td>
<td>0.3</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.83</td>
</tr>
<tr>
<td>Q12</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>0.5</td>
</tr>
<tr>
<td>Mean</td>
<td>0.79</td>
<td>0.80</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A1.5: Accuracy of ratings for Experiments articles

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q5</td>
<td>No</td>
<td>0.38</td>
<td>Yes</td>
<td>0.64</td>
<td>Yes</td>
<td>0.83</td>
</tr>
<tr>
<td>Q6</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>0.36</td>
<td>No</td>
<td>0.33</td>
</tr>
<tr>
<td>Q7</td>
<td>Yes</td>
<td>1</td>
<td>No</td>
<td>0.36</td>
<td>No</td>
<td>0.08</td>
</tr>
<tr>
<td>Q8</td>
<td>Yes</td>
<td>0.85</td>
<td>Yes</td>
<td>0.73</td>
<td>No</td>
<td>0.33</td>
</tr>
<tr>
<td>Q9</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.64</td>
<td>No</td>
<td>0.33</td>
</tr>
<tr>
<td>Q10</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>0.58</td>
</tr>
<tr>
<td>Q11</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.55</td>
<td>No</td>
<td>0.25</td>
</tr>
<tr>
<td>Q12</td>
<td>Yes</td>
<td>0.77</td>
<td>Yes</td>
<td>0.82</td>
<td>No</td>
<td>0.50</td>
</tr>
<tr>
<td>Q13</td>
<td>Yes</td>
<td>0.85</td>
<td>Yes</td>
<td>0.91</td>
<td>Yes</td>
<td>0.83</td>
</tr>
<tr>
<td>Q14</td>
<td>No</td>
<td>0.38</td>
<td>No</td>
<td>0.55</td>
<td>No</td>
<td>0.25</td>
</tr>
<tr>
<td>Q15</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.45</td>
<td>No</td>
<td>0.33</td>
</tr>
<tr>
<td>Q16</td>
<td>Yes</td>
<td>0.85</td>
<td>N/A</td>
<td>0.55</td>
<td>N/A</td>
<td>0.25</td>
</tr>
<tr>
<td>Q17</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.91</td>
<td>Yes</td>
<td>0.92</td>
</tr>
<tr>
<td>Q18</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.91</td>
<td>No</td>
<td>0.08</td>
</tr>
<tr>
<td>Q19</td>
<td>Yes</td>
<td>0.38</td>
<td>Yes</td>
<td>0.55</td>
<td>No</td>
<td>0.67</td>
</tr>
<tr>
<td>Q20</td>
<td>Yes</td>
<td>0.92</td>
<td>Yes</td>
<td>0.73</td>
<td>No</td>
<td>0.08</td>
</tr>
<tr>
<td>Q21</td>
<td>Yes</td>
<td>0.85</td>
<td>Yes</td>
<td>0.73</td>
<td>No</td>
<td>0.25</td>
</tr>
<tr>
<td>Q22</td>
<td>Yes</td>
<td>0.85</td>
<td>Yes</td>
<td>0.73</td>
<td>No</td>
<td>0.08</td>
</tr>
<tr>
<td>Q23</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>0.83</td>
</tr>
<tr>
<td>Mean</td>
<td>0.83</td>
<td>0.69</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

in absolute agreement with the gold standard. The mean accuracy for each group is shown at the bottom of the table. The gray lines at tables A1.4 and A1.5 group the checklist items according to the research phases classification (see Section 5.3.4).

It is also important to note that Q4 of the checklist for assessing an experiment is an open question: *What is the population of the experiment*. In that sense, it was not possible to compute quantitative values, such as accuracy and consistency.

In general, the percentage of students’ ratings that agree with the gold standard is good (0.74 for case studies, 0.65 for experiments). The results are not consistent across the studies though, and the ones presenting lower quality (i.e., C and Z) showed overall less accuracy.
Teaching students critical appraisal of scientific literature using checklists

Figure 5.2 illustrates the distribution of students’ ratings according to each study. The vertical bars show that the distribution of students’ ratings is similar to the gold standard. An exception occurs in article Z, in which students assessed the quality of the study substantially more positive than the gold standard.

RQ2. Consistency of Students’ Assessment

Consistency among the students’ assessment is given by degree of concordant ratings. As an example, 100% of participants that assess the case study A rated Q1 = yes; thus the rates are absolutely consistent (see Table A1.4). A conflicting assessment (i.e., low consistency) is produced when ratings are more equally distributed among all the possible options.
5.4 Results

To visualize such consistency, we generated stacked bar plots for each article i.e., case studies A, B and C, and experiments X, Y and Z (see Figure 5.3). Each horizontal bar can be divided in up to three different colors, representing the percentages of ratings that contribute to the total. Good consistency is depicted by the more monochromatic bars, while segmentation shows poor levels of consistency. Note also that the questions can be grouped according to the research phases (see Tables A1.2).

Case Study Checklist

We noted a few cases of absolute concordance on the students’ ratings of the case study checklist (see Figure 5.3). However, none of these are observed on the same checklist item across all the groups. Thus it is likely due to relevant information being available in the specific articles.

At large, the participants seem to rate more consistently the items related to (P2) design and (P3) operation phases (see Section 5.3.4). Conflicting rates are more frequent in relation to the (P4) data analysis phase. Finally, the consistency is gradually higher for studies A, B, and C, respectively.

Experiment Checklist

Overall, students’ accuracy using the checklist of experiments is similar to that of the case study checklist. On four items - Q13, Q17, Q18 and Q23 - we see a good degree of consistency across the three groups (see Figure 5.3). In most of these cases, the rates are also very accurate with the gold standard, except for Q18 rates of group Z.

Regarding the research phases (see Section 5.3.4), (P3) operation and data collection are the less consistent ratings. Good consistency is obtained on 3 out of 4 checklist items related to (P4) data analysis. Ratings related to (P5) presentation and reporting are also moderately consistent. The overall consistency on the rates of study X is higher than the other two studies.

RQ3. Students’ Perception of the Checklists

The students were also asked to provide their perception regarding three distinct aspects of using the checklists: i) ease of use; ii) understandability; and iii) ease of answer (see Appendix 5.8.3). This is primarily a reflection on the use of the checklists and secondly on the difficulty that the student perceived when assessing the examples of scientific literature.
Teaching students critical appraisal of scientific literature using checklists

Figure 5.3: Consistency in-between groups of the checklist for Case Study Research (top) and Experiment (bottom). Each horizontal bar shows the distribution of students’ ratings for a particular question.
Figure 5.4: Students’ assessment on ease of use, understandability and ease of answer of the checklists for Case Study (top) and Experiment (bottom)

The three aspects were rated on a Likert scale ranging from Strongly Disagree to Strongly Agree. Overall, the students’ perception is positive for ease of use and understandability of both case study and experiment checklists, as shown in Figure 5.4. Despite that, some responses pointed improvements needed, such as assessing the scope and context of the experiment, and the structure of the report.

**Students’ Perceptions of the Checklist Items**

The participants were also asked to identify the checklist items that were difficult to understand or hard to answer, and what made them hard to answer. Circa 64% of them participated, often enumerating more than one item. Figure 5.5 illustrate the frequency that each item was mentioned.

Regarding the case study checklist, the participants identified Q7, Q9, Q10, and Q11 as both hard to understand and to answer. These items are related to (Q7) repeatability and transparency of analysis procedures; (Q9) analyzing threats to validity; (Q10) triangulation of data or methods; and (Q11) address-
Teaching students critical appraisal of scientific literature using checklists

Figure 5.5: Items of the checklists for Case Study (top) and Experiment (bottom) that are hard to understand and hard to answer.

...ing ethical issues. Moreover, three of these items (Q7, Q9, and Q10) are related to the (P4) data analysis phase.

Overall, students mentioned that the difficulties to answer are related to the article itself i.e., not enough information is provided. Some also mentioned checklist-related issues, such as ambiguity, terminology and structure problems.

In relation to the experiment checklist, items Q16, Q15 and Q22 were respectively perceived as the more difficult to understand and to answer. They cover aspects related to the (P4) data analysis, such as (Q16) significance level;
(Q15) statistical tests; and (Q22) meta-analysis. Around 22% of participants also reported difficulty to understand item (Q12) Is data provided?

According to the participants, their own prior knowledge and experience, particularly with statistical analysis, is the factor that makes these items hard to answer. Some participants also mentioned checklist issues related to the subjectivity, and difficulty to justify the N/A ratings.

5.5 Discussion

This section discusses the findings of our evaluation study from the perspective of research and education.

5.5.1 Interpretation of the Results

The results of the evaluation draw our attention to a series of insights as follows:

The students performed better on the articles that presented higher assessed quality (i.e., A and X)

This observation (related to Contributions C1 and C2) is true both on accuracy with the gold standard and the consistency within the groups. One can assume that is easier for novices to assess empirical studies if the proper information is provided in the article. This is also likely to affect mature reviewers. Although this insight is fairly intuitive, it suggests that the research community benefits from better-reported studies to produce more fair assessments. As highlighted by Wohlin [552], it is important to think about reporting such that it facilitates synthesis. Similarity, the findings indicate that reporting is also important to enable quality assessment.

The students experienced more difficulty in assessing the checklist items related to (P4) data analysis.

We noted a large variation in the ratings of accuracy and consistency in the checklist items related to this phase. Moreover, the participants mentioned that those questions are hard to answer, particularly regarding lack of experience and prior knowledge (Section RQ3.). Our study was conducted in the context of a research methodology course, just before the lecture covering data analysis (Section 5.3.2). We, therefore, assume that this is likely to be related.
Teaching students critical appraisal of scientific literature using checklists

The students found checklist items difficult to answer due to the presentation of information in the paper or their lack of knowledge.

This observation is mostly supported by the students’ reflection (Contribution C3) but also reflects poor accuracy and consistency rates identified in those related checklist items. According to most of the student’s comments, the checklist itself is not difficult to answer, but it relies on the structure and contents of the paper and their capabilities to assess it.

5.5.2 Implications for Pedagogical Practice

The students’ perception regarding the critical appraisal activity was overall positive. The activity required them to reflect on the quality of the articles, and the application of the assessment checklists. We believe that this pedagogical approach effectively promotes critical thinking and analytical skills [549].

We here list our insights of this evaluation study and improvements suggested in our research methodology course. They are derived from our pedagogical reflection regarding the process employed in this study and the main findings (see Section 5.5.1). These suggestions are likely to benefit the pedagogical practice of empirical software engineering (Contribution C4):

- Issues to understand the checklist items related to lack of instruction (see Section 5.5.1). To better validate the students’ understanding of the whole research process, the critical appraisal activity should be conducted after they are provided lectures covering all the related research phases.

- Subjective criteria (e.g., are the data sufficient?) of the checklists are difficult to assess. There is a need to clarify the more subjective aspects by providing instructions on how to apply the checklists.

- Difficulties to assess the threats to validity (see Section RQ3.). The guidelines provided to the students do not cover the countermeasures to reduce validity threats in case study research. We should provide students additional supporting literature on the topic.

- Presenting both good and bad examples can help to demonstrate to students the best practice to adopt and practices to avoid. Our choice of objects (see Section 5.3.3) ensured that all students received a combination of different quality articles. To increase the likelihood of achieving this goal, we further suggest the teacher to follow-up and discuss the critical appraisal with the class.
5.5 Discussion

5.5.3 Validity threats and limitations

Here, we discuss a series of issues that may threaten our findings and the actions we took to minimize them.

Construct validity

We considered a set of factors (i.e., accuracy, consistency and reviewers’ perception) intended to properly address our research questions. However, we acknowledge that this is not the only possible set of constructs and that confounding factors (e.g., prior knowledge of the students) are likely to affect outcomes.

We intentionally removed three items from the checklist for reviewing experiments presented to the students. This was done to adjust the assignment to the students’ knowledge. As a result, our study does not investigate the scope and context aspects of the mentioned checklist (see Section 5.3.3).

Conclusion validity

The gold standard (Section 5.3.6) is used as a reference for comparing the students’ assessment with experienced researchers. As the checklist items are qualitative in nature and have an abstraction that leaves room for interpretation, a Yes as well as a No answer to a question may be reasonable, depending on the interpretation. For example, in Q2 for case studies - Are the case and its units of analysis well defined? one may debate what constitutes “well defined”. Through the use of experienced researchers independently assessing and after that discussing the papers, the threat of subjectivity in the ratings in the gold-standard could be reduced, but not completely ruled out given the nature of the checklist. A similar procedure of inter-rater agreement is, for example, followed in study selection processes for systematic reviews and mapping studies [500,526,538].

Internal validity

Unlike a real-world situation where a reviewer will likely work alone when applying the checklist, the students worked in pairs. It would have given us more data points by assigning students to work alone, thus making the research results more reliable. However, we compromised the quantity of data to provide the students a better opportunity to reflect and learn from each other when working in pairs. Another limitation is that we could not control the students
Teaching students critical appraisal of scientific literature using checklists
discussed their assessment outside their assigned groups or if they used sources beyond just the given checklist and guidelines.

As part of the assignment students were explicitly asked to read the following guidelines for conducting and reporting experiments [553] and case study [542] research in SE. Therefore, students may have additional information about critical aspects of reporting fresh in their minds beyond just the checklists. Furthermore, these guidelines may provide the implied knowledge and background necessary for using the checklists.

Another trade-off we made between the research and pedagogical objectives was when we decided to take the combinations on the diagonal of Table A1.3 from the possible treatments. We believe the students will have a better opportunity to learn if they apply checklists on articles with varied quality, whereas the strength of the checklist is tested well not on extremely good or bad examples.

**External validity**
The participants of this study are students within the context of master’s course on research methodology (as described in Section 5.3.2). We cannot assume that they accurately represent the population of novice researchers. However, the course is intended to provide students the skills needed for a researcher e.g., theoretical knowledge on research methods and the ability to critically evaluate scientific literature. The students have trained in various ways. During the first six courses, they have been reading and using scientific literature needed to solve their assignments.

Furthermore, training in research methods has been taking place in the course. Falesi et al. [510] debated the use of students as well as professionals in experiments. They, for example, highlighted the suitability of improving experimental designs (in our study the checklists itself). Furthermore, treatment conformance with students may be higher than with professionals. Hence, using students has its strengths, but also its weaknesses (such as lower generalizability to experienced reviewers).

We took several steps to ensure that students’ grades are not affected by this study and that the results of this study are not biased by our impression of the students. We anonymized students’ answers (by removing the names of the students and only keeping the Group-ID to track assignment of treatments and objects). Furthermore, the results (in this paper) have been aggregated and cannot be traced back to individual students.
5.6 Conclusions

To ensure that the grades of the students in the course are not affected by the study we conducted an independent analysis of the student answers. That is the data analysis was conducted independently and by a different researcher than the one who participated in the collection of data and grading of the students’ assignment.

5.6 Conclusions

In this paper, we reported an evaluation exercise conducted with Master’s student to investigate the support of checklists for appraising empirical studies. Our findings could be summarized as follows: 1) students’ accuracy differs when compared to more mature researchers; 2) the students’ ratings are not particularly consistent; and 3) the students acknowledge the benefits on using such instruments, but mentioned several checklist items that are hard to answer.

From the pedagogical perspective, the appraisal activity is intended to promote critical thinking in the students. This is likely to further impact the research practice, as students transition to mature researchers. We believe that this entwined relation between teaching and research is beneficial to all, and our insights are intended to contribute to both communities.

From the research perspective, the students’ perception provides us insights on the quality of the checklists. We identified issues related to “ease of understanding” and “ease of answering”, as well insights on how to apply and to evolve the checklists. We outlined the main limitations of our study design and trade-offs we made to align this investigation with the course objectives.

In future work, we plan to investigate the application of checklists by experienced researchers. By comparing their results with the novices, we can highlight which checklist items benefit more from experience. Issues that are similar for both groups are likely to be related to the checklist instrument or its underlying theory, thus important to be further explored by the research community.

5.7 References


Teaching students critical appraisal of scientific literature using checklists


5.7 References


Teaching students critical appraisal of scientific literature using checklists


5.7 References


Teaching students critical appraisal of scientific literature using checklists


5.8 Appendix

5.8.1 Checklist for Reading and Reviewing Case Study Research [542]

The reader of a case study report - independently of whether the intention is to use the findings or to review it for inclusion in a journal - must judge the quality of the study based on the written material. Case study reports tend to be large, firstly since case studies often are based on qualitative data, and hence the data cannot be presented in a condensed form, like quantitative data may be in tables, diagrams, and statistics. Secondly, the conclusions in qualitative analyses are not based on the statistical significance which can be interpreted in terms of a probability for an erroneous conclusion, but on reasoning and linking of observations to conclusions.

Q1) Are the objective, research questions, and hypotheses (if applicable) clear and relevant?

Q2) Are the case and its units of analysis well defined?

Q3) Is the suitability of the case to address the research questions clearly motivated?

Q4) Is the case study based on theory or linked to existing literature?

Q5) Are the data collection procedures sufficient for the purpose of the case study (data sources, collection, validation)?

Q6) Is sufficient raw data presented to provide understanding of the case and the analysis?

Q7) Are the analysis procedures sufficient for the purpose of the case study (repeatable, transparent)?

Q8) Is a clear chain of evidence established from observations to conclusions?

Q9) Are threats to validity analyses conducted in a systematic way and are countermeasures taken to reduce threats?

Q10) Is triangulation applied (multiple collection and analysis methods, multiple authors, multiple theories)?
Q11) Are ethical issues properly addressed (personal intentions, integrity, confidentiality, consent, review board approval)?

Q12) Are conclusions, implications for practice and future research, suitably reported for its audience?

5.8.2 Checklist for Reviewing Experiments [553]

Below is a list of questions, which are important to consider when reading or reviewing an article presenting an experiment. The list below should be seen as a checklist in addition to normal questions when reading an article. Some specific aspects to consider when reading an experiment article are:

Q1) Is the experiment understandable and interesting in general?
Q2) Does the experiment have any practical value?
Q3) Are other experiments addressing the problem summarized and referenced?
Q4) What is the population in the experiment?
Q5) Is the sample used representative of the population?
Q6) Are the dependent and independent variables clearly defined?
Q7) Are the hypotheses clearly formulated?
Q8) Is the type of design clearly stated?
Q9) Is the design correct?
Q10) Is the instrumentation described properly?
Q11) Is the validity of the experiment treated carefully and convincing?
Q12) Are different types of validity threats addressed properly?
Q13) Has the data been validated?
Q14) Is the statistical power sufficient, are there enough subjects in the experiment?
Q15) Are the appropriate statistical tests applied? Are Parametric or non-parametric tests used and are they used correctly?
Q16) Is the significance level used appropriate?
Q17) Is the data interpreted correctly?
Q18) Are the conclusions correct?
Q19) Are the results not overstated?
Q20) Is it possible to replicate the study?
Q21) Is data provided?
Q22) Is it possible to use the results for performing a meta-analysis?
Q23) Is further work and experimentation in the area outlined?

5.8.3 Additional Questions Regarding the Perception of Use of the Checklists

Please answer the following questions regarding the use of the given checklist:

A) Please respond to what extent do you agree or disagree with the following statement: Overall, the checklist was easy to use.
   (strongly agree | agree | neutral | disagree | strongly disagree)

B) Please respond to what extent do you agree or disagree with the following statement: The questions were easy to understand.
   (strongly agree | agree | neutral | disagree | strongly disagree)

C) Please write the question numbers (from Table 1), if any, that were difficult to understand.

D) Please respond to what extent do you agree or disagree with the following statement: The questions were easy to answer.
   (strongly agree | agree | neutral | disagree | strongly disagree)

E) Please write the question numbers (from Table 1), if any, that were difficult to answer.

F) What made it difficult to answer these questions?

G) Does the checklist cover all the important aspects (as mentioned in the guidelines for conducting case study research) for high quality case study research?
Chapter 6

An Empirically Evaluated Checklist for Surveys in Software Engineering

This chapter is based on the following paper:


6.1 Introduction

A survey is a widely deployed research method in the area of Software Engineering (SE) and an increase in its usage has been highlighted by, e.g., Punter et al. [599]. Its purpose is to investigate a population, in order to construct explanatory models [556,609] or to validate knowledge [581,586]. Survey research is often employed when there is a need to study a large set of variables [609] or to perform a retrospective analysis [597]. It may be used to draw conclusions based on both quantitative and qualitative data [564].

Researchers have highlighted various challenges during the survey process. Common challenges are the formulation of questions [605], so to avoid shortcomings (e.g., introducing bias inside questions [609]), and the identification of invalid responses [610]. Other challenges are related to the recruitment of
participants, such as how to obtain a sufficient number of responses and how to prevent high drop-out rates [554,576].

The need for improving the standards of conducting and reporting survey-based research, in particular with respect to the definition of the population and sampling strategies is evidenced by Stavru [603]. Furthermore, Stavru pointed out a lack of checklists for auditing surveys in SE which could be of help to both researchers conducting survey research as well as to those evaluating and reviewing the research.

Motivated by these needs, we employed an empirical approach to constructing and evaluating an checklist\(^1\) for survey-based research in SE. Our empirical approach comprises of two steps (see Sections 6.3 and 6.4):

First, we detail the process to construct a checklist grounded on survey guidelines. The method used to derive the checklist was guided by two principles: (a) identify existing guidelines for survey research; in the context of SE, 12 methodological studies have been considered; (b) elicit the process stages, recommended practices, and related rationales. Those rationales support the cost-effectiveness analysis of employing a set of related practices.

The method for systematically deriving the checklist was based on thematic analysis [563]. Vote counting was applied to the themes identified in order to compute the frequency in which they occurred. Further, a co-occurrence was obtained through a relationship matrix relating different categories (e.g., practices versus rationales, and practices versus rationales).

Later, we evaluated the checklist in the context of reviewing survey reports by using a mixed research approach [581]. The evaluation process involved two distinct phases: (a) to apply the checklist on a set of published survey reports and register the assessment scores, and (b) to verify the results of this assessment with the corresponding authors of those survey reports.

The assessment produced a compliance coefficient for the selected studies in relation to each of the checklist items. We further investigated the authors’ feedback in order to understand patterns we identified in the assessment scores. We also collected and addressed suggestions from the experts to improve the checklist instrument.

The remainder of the paper is structured as follows: Section 6.2 describes the background and related work. Section 6.3 details the systematic approach we used to construct the checklist. The evaluation of the checklist in research practice context is presented in Section 6.4. Section 6.5 discusses the findings and finally, and Section 6.6 concludes the paper.

---

\(^1\) The resulting instrument is further detailed in Appendix 6.8.3.
6.2 Background & Related Work

We first present existing survey guidelines that are subject-independent or that have been proposed in other fields. Thereafter, an overview of SE specific guidelines is provided. After giving an overview of the guidelines we describe the literature on survey assessment. Finally, we looked into checklists proposed to support Empirical Software Engineering (ESE).

6.2.1 Survey process and guidelines

Survey as a research method has been established in social research for half a century. It has been employed in several academic fields, such as health care, politics, psychology, and sociology [562]. As a consequence, methodological knowledge on surveys has been published first in these fields.

As the survey research method matured, cross-field guidelines appeared (e.g., [555, 573, 574]). These publications aimed to provide methodological support independent from the subject of research. Nevertheless, it is not uncommon for their mentioned practices to focus on the social aspects of research.

Those guidelines [555, 573, 574] describe a survey-research process comprising a set of stages, such as question design, sampling, data collection, instrument evaluation, measuring and data analysis [574]. Survey research is acknowledged for being flexible, although the process stages are often conducted sequentially. It is worth mentioning that the survey process is a complete research method (i.e., including planning, execution, analysis, and reporting); the survey data collection instrument is called a questionnaire.

In addition to describing the process, the guidelines also recommend best practices based on desirable attributes for high-quality surveys. Such quality is based on evaluation of the produced evidence (e.g., precision, credibility), ethical issues (e.g., consent, privacy) and mitigating validity threats (e.g., sample error, non-responses) [574, 589].

6.2.2 Survey guidelines in Software Engineering

The need for specific and tailored guidelines to conduct empirical research in the context of SE has been pointed out, e.g. [594, 602, 608]. This demand is especially relevant to formal experiments and case studies, due to the popularity of such methods, but also applies to survey-based research. Methodological support for surveys in SE first appeared around the 1990s [593].
Three main guidelines [580, 586, 590] detail the survey research process in the SE field. They jointly provide a comprehensive structure for the research process, despite differing slightly from each other. Major differences are in relation to the breakdown structure of process stages and the recommended practices provided.

Besides these three main publications, a series of additional studies extend the guidance to particular stages of the survey process. For example, the challenges of identifying the target audience and establishing a sampling frame are discussed in [565–568, 570]. The recommended practices in this set of papers are complementary, although some partially overlap.

Other studies provide lessons learned from carrying out the process in different contexts:

- Punter et al. [599] focus on self-administered online surveys and address issues such as monitoring real-time responses, identifying the reasons for dropouts and encouraging participants to complete a survey instrument;
- Ciolkowski et al. [558] addresses practical issues related to the process itself, such as managing resources and ensuring that deviations do not threaten the completion of the entire process; and
- Cater et al. [557] address replication challenges, such as updating a survey instrument, collecting data and comparing the results.
- Conradi et al. [560] and Ji et al. [579] provide challenges and lessons learned from applying survey research in international contexts, especially regarding invitations and managing responses.
- Torchiano et al. [604] provide lessons learned from experiences in conducting many survey studies, covering topics such as the study design, identifying target audience, sampling, questionnaire design and managing responses.

Additional references for survey-based research are provided in our previous works [592, 593].

### 6.2.3 Survey assessment

When reviewing existing guidelines (see Section 6.2.2) we found out that several researchers highlighted the need for an instrument to audit survey research in SE context. This need is further stressed by the lack of reporting of the employed criteria to assess survey research [603].
Stavru’s work [603] provides a critical review of surveys in the area of agile software development. In order to carry out the review, Stavru used 21 criteria by which the thoroughness in reporting surveys was assessed. These criteria were extracted from different sources, cf. [558,580,586,591,598,601]. Note that the method of eliciting the criteria was not detailed.

Stavru also highlighted that the different criteria were not equally important, and rated them on a scale from one to five. The most important criteria that ought to be documented were:

- Sampling frame, method, and size
- Response rate
- Assessment of a survey’s trustworthiness
- Survey process
- Conceptual model comprising of the constructs investigated (e.g., variables and their relations)
- Target population
- Questionnaire design

### 6.2.4 Checklists in Software Engineering

Checklists have been proposed for various research methods with a specific focus on their usage in the SE context. Looking at the ways in which checklists were built, researchers most often based the construction of a new checklist upon existing ones (cf. [577,607]).

As an example, Kitchenham et al. [583] combined two checklists [572,584] to assess experiments and to evaluate whether researchers may use them objectively. Their findings indicate that a larger number of reviewers was needed (eight) to reliably assess studies using their checklist, which could be improved by having researchers conduct reviews in pairs (cf. [583]). Additional checklists proposed for assessing experiments are, e.g., [578,583,609].

Höst and Runeson [577] put forward a checklist for case study research, divided according to the research stages, including design, preparation for data & evidence gathering, data analysis, as well as reporting. A validation of the checklist identified that it is appropriated to support researchers conducting case studies, but it is too extensive for reviewing case study reports.

Based on the validation results, Höst and Runeson [577] also created a reduced reviewer’s checklist abstracting the original checklist to reduce the number
of items to be checked. The 38 items from the researcher’s checklist were condensed into 12, most of which synthesizing a list of practices suggested in the original checklist. Additional checklists for reviewing case study research in SE are found in, e.g., [582, 595, 609].

Wieringa [606] observed that the individual checklists with the same focus differed, which may result in confusions for reviewers. The author highlights the need to find common checklist items across research types as they may share specific aspects. Thus, Wieringa et al. [607] used existing checklists (e.g., [577, 578, 600]) for experiments and case studies as a basis to synthesize an unified checklist. Later, the authors evaluated their checklist by having them used by PhD students and researchers in different research groups, as well as by conference participants.

Stavru’s [603] filled a gap in the existing body of knowledge by complementing the set of available checklists with a set of criteria for assessing survey research. No other checklists to assess surveys were identified in our systematic literature search (cf. [592]).

Great emphasis was placed upon (a) basing the checklist on existing literature, and (b) following a systematic approach to eliciting checklist items [603, 606]. Thus, our work complements the above-mentioned by deriving and evaluating an assessment checklist grounded in existing guidelines for survey research.

6.3 Step 1. Construction of the checklist

The first step of our research approach entailed the systematic construction of the checklist. Three sub-contributions are made that ultimately lead to the checklist proposed:

C1. **Consolidation of survey processes and decision points:** We present a consolidated survey process based on existing guidelines. Key decisions points and implications of decision-making are highlighted. For example, a key decision in a survey process is the type of sampling used, which impacts participant recruitment and data analysis. Our checklist has to be adapted depending on the decisions taken.

C2. **Extraction of recommended practices and their mapping to the survey process:** We extracted the recommended practices to be carried out during a survey research process, which were later mapped to the research process identified in C1. Mapping the practices to the main stages aids researchers
in the planning of surveys, as it indicates in which process step a practice is executed and where its impact needs to be considered.

C3. *Extraction of rationales for the recommended practices*: The reasons for considering existing survey research practices should be motivated by a rationale, thus making the value of adopting such practices explicit. This is particularly pressing because a survey’s cost-effectiveness is an important consideration. Thus, understanding the rationales for the recommended practices supports the cost analysis of a practice and its effectiveness (i.e., the rationale regarding the value a given practice adds to the survey research).

### 6.3.1 Method

**Research questions**

We formulated three research questions corresponding respectively to each of the three contributions stated above, as follows:

RQ1. Which stages and key decisions are specified for the survey process (C1)?

RQ2. Which practices are suggested and how do they map to the stages of the research process (C2)?

RQ3. What is the rationale for conducting the respective recommended practices (C3)?

**Study identification and selection**

In order to select an appropriate set of primary studies, we used evidence from a systematic mapping study we conducted in 2016 [592, 593]. Initially, we have gathered papers from major conferences and journals related to ESE:

- **EASE** International Conference on Evaluation and Assessment in Software Engineering (1997-2014)
We employed a manual search and snowballing approach to identify methodological papers for different research methods in SE, including survey-based research. The selection resulted in 341 papers overall [593], 39 from which provide methodological support for surveys in SE research.

In order to address the research questions of this study, we included 12 studies from this set, as they provide guidelines and recommendations for survey-based research. Papers that provide recommendations for a specific context only (e.g. Global Software Engineering [560,579]) and papers that describe methodological support other than guidelines and recommendations (e.g. classification schemes and evaluation instruments) were excluded. A detailed discussion on the contents of each included study is given in [592].

We categorized the included studies according to the survey process activities they provide support, and the support provided (i.e., comprehensive guidelines, or issues, recommendations and lessons learned), as presented in Table A1.1. The process phases we use here are derived from the process described in the three main guidelines [580,586,590].

The first category, i.e. Guidelines for Survey Research, includes three papers that cover the overall survey research process. One of these papers, i.e. Principles of Survey research [586] is actually divided into 6 parts, each of which published individually. The six additional entries under Survey Issues, Recommendations and Lessons Learned address specific activities of survey research such as sampling, instrument design, and validation, recruitment and response management. The sampling strategies series [566–569] is a set of four papers that investigate large scale sampling issues in surveys.

Data extraction and analysis

From the set of of methodological papers in Table A1.1 we extracted data to address the research questions RQ1 to RQ3. We employed a thematic analysis process [563] to identify and analyze themes related to three major categories:

- Process stages, e.g., data analysis.
- Recommended practices, e.g., identify reasons for non-responses.
- Rationale attributes, e.g., representativeness.

The thematic analysis process followed the framework proposed by [563], as follows:
6.3 Step 1. Construction of the checklist

Table A1.1: List of included papers and related survey process phases and activities covered.

<table>
<thead>
<tr>
<th>Planning</th>
<th>Execution</th>
<th>Replicating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research objectives</td>
<td>Design instrument</td>
<td>Analysing results</td>
</tr>
<tr>
<td>Target audience</td>
<td>Evaluate instrument</td>
<td>Reporting results</td>
</tr>
<tr>
<td>Sampling plan</td>
<td>Manage responses</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Guidelines for Survey Research

| Principles of Survey Research, pt. 1-6 [586] | · · · | · · · | · |
| Designing an Effective Survey [580]         | · · · | · · · | · |
| Guidelines for Surveys v. 1.1 [590]         | · · · | ·     | · |

Survey Issues, Recommendations and Lessons Learned

| Conducting On-line Surveys [599] | · · | · · | · |
| Practical Experiences [558]      | · · | · · | · |
| Personal Opinion Surveys [558]   | · · | · · | · |
| Surveys as Secondary Studies [588]| · · | · · | |
| Sampling Strategies [566-569]    | ·   | ·   | |
| Replication of Surveys [557]     |     |     | · |

1. **Extract data.** We collected the included studies and aggregated them in a common list using Atlas.ti [575] - a qualitative data analysis software. We also collected bibliographic information for further reference to the included studies.

2. **Code data.** We read all the included studies, identifying segments of the text related to the three main categories and associating them with codes. The segments are characterized by a level of granularity of one paragraph, notwithstanding a single paragraph is often associated with more than one code. Paragraphs containing no relevant information were associated with no code.

3. **Translate codes into themes.** The terminology for initial codes is derived from the three guidelines analyzed first (i.e., [580, 586, 590]). Later,
we iteratively improved and updated the initial codes according to the different views presented by the additional sources. Successive iterations of the process further refined the theme set, by combining or merging synonyms and aggregating related themes into families (e.g., representativeness is part of the external validity rationale). We also removed duplicates and combined successive occurrences of the same theme in larger segments, thus comprising several paragraphs.

4. **Create a model of higher-order themes.** Later, we identified and analyzed the relationship between themes within the three categories. In particular, we made explicit the recommended practices more frequently related to each process stage and the rationale to adopt them. We computed a co-occurrence coefficient [561] to analyze how frequently two related terms occurred alongside each other. Models describing such relationship are given in Sections 6.3.3 and 6.3.3.

5. **Assess the trustworthiness of the synthesis.** Finally, the identified process stages, recommended practices and rationale are used to answer the research questions. In addition to that, we constructed a checklist instrument grounded on the themes and the relationship among them. We assume that this comprehensive checklist can assess survey research about the relevant practices proposed in the guidelines for SE. Evidence collected from the literature supports this assumption, and we further evaluate the checklist with research practitioners (see Section 6.4).

### 6.3.2 Threats to validity

**Construct validity.** To ensure a similar understanding and reduce research bias on the thematic analysis, we piloted the coding process between the three authors. The results imply a fair agreement, i.e., in average, 46.5% of the themes are similar, although worded differently. Further, based on the reflections of the pilot study, the first author coded the remaining papers. The co-occurrence coefficient used in the analysis takes into consideration the position and size of sentences but is prone to non-significant values when comparing themes that differ largely in size [575]. We partially address this potential bias by normalizing the values within the same row.

**Internal validity.** Internal validity relates to factors affecting the outcome of the study not accounted for by the researchers. One threat is the bias in interpreting the findings. Hence, at each stage of the research, the intermediate results were discussed among the researchers (observer triangulation).
Our data set consists of 12 studies gathered from a previous literature study [592]. We relied mostly on the original study design to the search and selection stages, using its reported evidence to collect our relevant set. We employed structured reading and coding to analyze the data set, producing themes and higher-level categories. The first author conducted the data extraction and analysis, further discussing the resulting themes with the other co-authors. We trust that this iterative process minimized the judgment bias of a solo researcher.

**Conclusion validity.** One threat to conclusion validity is whether the data based on which the survey was created is complete. The additional 6 papers in our selection (see Section 6.3.1) complemented the study results with 10 extra practices and 1 rationale. Those extra themes are related to particular challenges a researcher may face during the process, namely a large sample frame [566, 567], managing online surveys [599], and survey replication [557]. By adding guidelines very specific to the individual stages of survey research increased confidence in the results, despite being limited by the availability of the literature for each stage.

**External validity.** The resulting themes and frequencies were extracted from relevant methodological guidance for SE. However, we cannot assume that the practices and rationales identified are only important for this field. Moreover, there is the possibility of identifying a valid theme outside of our data set, e.g., a non-selected paper or the practical experiences of a researcher. We, therefore, conducted an evaluation of our proposed checklist with SE researchers, thus investigating its appropriateness and identifying potential improvements (see Section 6.4).

### 6.3.3 Results

The following section is structured according to the three contributions proposed in Section 6.3.1. Each section is, in turn, broken down into its units of analysis (i.e., decision-making points, process stages, and rationale aggregation).

**C1) Survey Research Process**

Figure 6.1 presents an aggregation of the survey research processes described by Kitchenham & Pfleeger [586], Kasunic [580], and Linäker et al. [590]. Although the main stages (and terminology) slightly differ among the guidelines, the processes are similar and follow a sequential flow. We adopted the view from [586] to describe the execution phase comprising two stages, one for recruiting participants and one for administering responses.
We also identified key decision-making points during the process, two of which should be addressed during the sampling stage (i.e., D1 and D2) and two others during the instrument design stage (i.e., D3 and D4). Those conditional nodes require researchers to make decisions regarding a survey’s research design that can potentially impact the subsequent stages.
D1. **What kind of sample is selected?** Depending on the strategy for selecting respondents, the researcher can choose a **probabilistic** (e.g., random selection) or **non-probabilistic** (e.g., convenience, quota, snowballing) sample. This decision mainly affects the data analysis methods (as probabilistic samples are meant to be generalizable) and recruitment approaches (e.g., random selection) employed.

D2. **How are the participants recruited?** On the one hand, **self-selection** approaches allow for potential respondents to volunteer themselves which may introduce biases in the interpretation of the data. On the other hand, **personalized selection** (such as invitation letters and more rarely authorization codes) require specific actions for the recruitment and management of the responses. This decision-making point is often interdependent of the kind of sample (D1).

D3. **What type of survey instrument is designed?** **Self-administrated** surveys are mainly distributed in the form of Web pages or printed questionnaires, thus the respondents fill out the data themselves. **Interviewer-administrated** surveys include face to face or phone interviews where respondents provide the information to a researcher, who records the data. This decision not only drives the instrument design but can also heavily impact the execution stages (i.e., recruitment and response management).

D4. **What response formats are collected?** Question structure types could be **open-ended** and **close-ended**. Open-ended questions are less restrictive allowing for respondents to use their own words, whereas close-ended questions are represented by scales that can be easily quantified. This decision determines the data analysis methods employed, i.e., qualitative or quantitative approaches. Often survey instruments include a mix of both question types, thus requiring both analysis approaches.

**C2) Recommended Practices**

We identified a list of recommended practices for the survey research process and computed how frequently they occur alongside each other. A co-occurrence coefficient was calculated as follows: 

\[ c := \frac{n_{1,2}}{n_1 + n_2 - n_{1,2}} \]

where \( n_1 \) and \( n_2 \) are the vote-counting frequencies of two themes \( t_1 \) and \( t_2 \) respectively, and \( n_{1,2} \) is the joint frequency, i.e., how many times the two themes co-occur. An example of the coefficient computation is given in Figure 6.2.
An Empirically Evaluated Checklist for Surveys in Software Engineering

Figure 6.2: Example for computation of the co-occurrence coefficient, given that t1 occurs 30 times in the data set, t2 occurs 41 times, and they simultaneously occur 19 times.

The resulting relationships between recommended practices and process stages are presented in a co-occurrence matrix (see Table A1.4 in Appendix 6.8.1), where the cells are filled in teal-tones according to the coefficient value (see Section 6.3.3). Darker cells represent a stronger co-occurrence between two themes.

We opted for normalizing the themes in each matrix row since our analysis relates mainly to only comparing themes within the same category. Our normalized coefficient range from 0-100, whereas 100 relates to the maximum co-occurrence score in the corresponding group, and 0 corresponds to no co-occurrence. The non-normalized co-occurrence coefficients for each practice in relation to the process stages is available at http://bit.ly/2tRgW2t.

The matrix shows that each stage has a set of practices strongly associated with it, and also what other practices influence that stage. The influence is usually in the form of planning the practice to be carried out, follow-up actions, or consequences of a given decision regarding different practices to adopt (see Section 6.3.3).

A summary of the recommended practices associated to each process stage is given below:

1. Research objectives: Survey-based research is motivated by a specific goal. Thus it is important to state the research questions that correspond to such goal. The two main recommendations related to this stage are P1) to limit the scope, as this could impact upon the survey’s complexity, and P2) to apply the goal-question-metric (GQM) approach to define its objectives. Moreover, the questionnaire items and collected data should be mapped to the research questions (P44).
2. **Study plan:** The need for designing the survey research is set at the beginning of the process, often along with the research questions [580]. The main suggested practices for this stage are to: P3) investigate related work; P4) define a set of procedures to guide the process; P5) develop a schedule plan for the stakeholders; and P6) start a diary or log book. The study plan should then be iteratively revised during the process, and the updates recorded in the log book (P1). This information is specially required for the reporting stage, at the end of the process.

3. **Identify and characterize the population:** Audience analysis (P7) is often employed to identify and select the characteristics of the population addressed by the research. This task has a strong effect on the sampling stage, in which the sources of sampling (P10) should be defined. Surveys often target potential participants at open databases (P16), but could employ restricted databases (P10) as an alternative or complementary source of sampling. Restricted databases should be investigated prior to the sampling stage.

4. **Sampling plan:** It is often employed in order to sample the population representatively. A sample plan should contain the sources of sampling (P10), units of observation and search unit (P21). The type of sample (P8 and P20) should potentially lead the decision for the data analysis methods employed. Other essential aspects to be considered are the P11) size of the sample and P19) how to manage large samples.

Additional practices for this stage include to P14) remove the redundant units; P15) apply criteria for selecting the units of observation; P17) plan the retrieval of search units; and P22) partition the population according to the chosen characteristics. Strategies for recruitment (e.g., P8, P18, and P20) are likely to impact the participant selection stage.

5. **Instrument design:** A questionnaire or similar instrument is designed to gather data from the sample representative of the target population. Depending on the choice of distribution, the instruments can be P43) self-administered, e.g., online forms (P51), or P50) interview-administered e.g. interview or phone survey. They can be P29) prototyped, P30) implemented from the sketch, or acquired through P38) commercial tools or P31) reuse.

Several recommendations to design and present an instrument are provided in the literature, e.g., avoid P33) intrusive and unethical questions, and P57) to lead the respondents; provide P27) a progress indicator, P35) questionnaire navigation, P40) instructions of use, P41) option to resume answering,
and mainly P32) ask simple, unambiguous, actual and targeted questions. Responses can assume P46) open-ended or P49) close-ended formats.

6. **Instrument validation**: After design, the ability of the instrument to measure what is intended should be assessed. The most frequently cited approaches for the assessment are P66) piloting, P65) retest, P62) focus groups, and P63) expert or P58) non-expert reviews. Additionally, user-related metrics (e.g., usability, readability, time to respond) can result in improvements to the instrument design (P61). Ancillary documents supporting the recruitment stage should also be reviewed, e.g., cover letter (P60) and thank you letter (P56), likely providing incentives to the respondents (P67).

7. **Participant recruitment**: The strategies to select potential participants are previously defined in the sampling plan stage, such as P24) invitations and authorization codes, P26) self-recruitment, and P13) snowballing. By adopting proper actions and technology support, researchers can even investigate the potential threats to the process related to drop-outs (P68).

8. **Response management**: After distribution of the instrument to the selected participants, it is important to observe the response rate (P64) in order to identify the reasons for non-responses (P25). To ensure that the expected number of responses is achieved, researchers are likely to send reminders (P70) or to provide rewards for participation (P67).

9. **Data analysis**: Prior to the synthesis, the collected data should be validated (P78) in order to handle incomplete and missing values (P74). Furthermore, qualitative (e.g., P73) or quantitative (e.g., P76) analysis methods can be employed according to the survey’s sample and response format. The results should then be P80) presented, P81) interpreted and likely P77) compared to particular subsets of the population. An additional suggestion to ensure their reliability is P79) to have more than one item measuring the same variable.

10. **Reporting**: The main practice related to the reporting phase is to produce an output of the information contained in the process documentation (P6). Ideally, the documentation is to be updated during the survey process, including the data analysis and results’ interpretation. Both the related work (P3) and the adopted guidelines (P4) are used as additional information sources for this stage. Finally, it is important to consider the report’s intended audience (P7).
The frequency with which the practices occur in the segments of the text is not fully comparable to other practices, i.e., one can not judge a practice as more important if it is mentioned more often. However, the co-occurrence factor can be compared to other instances in the same row to identify in which process stage the practice is more relevant.

Moreover, some recommended practices are mutually exclusive (e.g. P8 probabilistic sample and P20 non-probabilistic sample) requiring the researcher a decision to adopt one of the alternatives. The reasons to adopt a particular practice over another depends on the researchers’ conscious decision supported by the guidelines employed.

C3) Rationales and Outcomes

We define a rationale as the motivation to choose a particular practice. They are often described as desirable process attributes (e.g., cost-effectiveness, generalizability) or outcomes of such actions (e.g., minimize or introduce bias). As an example, to achieve generalizability, a researcher should utilize probabilistic sampling (P8) and estimates of the population size (P23). Moreover, several of the rationales can be related to validity threats, i.e., concerns about the methodology in order to achieve valid conclusions [596]. The additional rationales are not fully linked to validity, but can still positively or negatively affect it (e.g., willingness is likely to influence data quality, and thus validity). The coding resulted in a set of nine rationale categories, as shown in Figure 6.3.

R1. Conclusion validity: the actual extent to which conclusions about the investigated relationship are true or correct. Survey-based research employing quantitative analysis methods are prone to significance, effect size, and magnitude factors. Moreover, the reliability or confidence of the results is inherent to the conclusion validity.

R2. Construct validity: refers to the interaction between the underlying theory and measurement constructs, i.e., if the variables are actually measuring what they mean to. The main rationale in this category is mensurability, mainly addressed by the instrument validation stage.

R3. External validity: the degree to which the results of the survey can be applicable to other scenarios, such as different contexts and strata of the population. Surveys are largely impacted by external validity factors, such as generalizability, replicability, and relevance to practice. Some major factors in this category are whether the sample is representative and heterogeneous to the overall target population.
R4. **Internal validity:** represents an estimate of the degree to which conclusions about the investigated relationships can be drawn based on the measures and the research process. In survey-based research, the rationales in this category are mostly related to the sampling and instrumentation stages, e.g., understandability and time to respond.

R5. **Interpretive validity:** related to the inference of the participants’ opinions from the collected responses. Unlike the conclusion validity, the interpretive is more focused on the analysis of the qualitative data. Thus, factors such as correctness, confirmability, and subjectiveness play an important role in interpreting the data.
R6. **Participant-targeted**: additional factors related to the respondents include concerns about anonymity and confidentiality, usability and willingness. Those factors are likely to impact the data quality, as they can positively or negatively influence the participants while answering the survey.

R7. **Process-targeted**: the improvement of the process itself is a target of several rationales. Researchers carrying out surveys should pay special attention to cost-effectiveness, as their decisions are likely to require extra resources. Other practical considerations include the complexity of the instruments and techniques, completeness of the sampling sources, and compatibility of produced data.

R8. **Researcher-targeted**: by providing their opinions, survey respondents trust that the gathered data will be processed responsibly. Thus, it is essential that the researchers are aware of potential ethical issues and their responsibilities regarding the survey process.

R9. **Result-targeted**: one would expect a properly conducted survey process to produce useful results. Data validation tasks are meant to assess data consistency and completeness. Moreover, the precision of the results can be achieved by properly addressing a representative sample of the target population.

Rationales are standards for quality of survey research. Moreover, by relating them to the practices can potentially support design decisions. To support such a decision-making process, we also computed a co-occurrence coefficient of the relationships between recommended practices and the rationale (available at https://bit.ly/2UpQqYf). This map makes explicit which practices to adopt to achieve particular standards.

While conducting the research process, it is essential to reflect on the importance of different rationales to be prioritized. As an example, to strengthen the external validity (R3), a researcher should take particular care of strategies for sampling and recruitment of participants, such as using a probabilistic sample (P8), updating the sampling frame based on additional collections from the same search unit (P12), and employing a snowballing technique (P13) to recruit participants.

**C4) Checklist Instrument**

To derive the checklist, we identified the relevant practices to each process stage according to the co-occurrence index. Further, we aggregated recommended
practices addressing a similar topic, e.g. the usability of an online questionnaire is addressed by the set of practices P27, P28, P35, P39, P40, P41, P45, and P51. We then looked up the segments of text coded according to these practices to derive a checklist item. In this way, the formulation of the checklist items closely reflects the practices suggested in the literature.

We also identified the rationale associated with the practices that originated a checklist item (see Section 6.3.3). Our intention by providing the rationale is twofold: (i) to better structure the checklist according to the reasons for adopting the set of practices, and (i) to support the use of the checklist with regard to decisions to be made for employing a set of related practices. Similarly, we identified key decision points that could impact the research (see Section 6.3.3, e.g., D1:P means kind of sample: probabilistic as D1:NP is related to the non-probabilistic alternative).

Later, the first author prepared a draft of the checklist instrument and discussed it together with the other authors. The draft was further refined based on the results from the discussions. Most of the refinements took the form of rephrasing and reordering the checklist items. A few checklist items were integrated to other to improve readability and understandability of the instrument (cf. [577]).

We also identified that most of the checklist items are related to the methodological rigor, and there is a gap about how to report the evidence produced by survey research. Thus, we complemented the reporting section of the checklist with more generic questions selected from the checklist proposed by Dybå and Dingsøyr [571]. Those questions are meant to assess the quality of the evidence produced by empirical studies, regardless of the research method employed.

The checklist resulting from this process had originally 38 items (see pre-evaluation checklist in Appendix 6.8.2). One can notice that several practices are presented within a stage not in accordance to the co-occurrence table (Section 6.3.3). Although some of the practices require early planning, they could eventually be carried out in a later stage. Therefore, the checklist is organized in accordance with the stages in which those actions are more likely to take place.

Several process stages and their recommended practices are subject to decision-making (e.g., the kind of sample, instrument type). The choice should be guided by the motivations (i.e., rationales) and desired outcomes of the process. Due to this decision-making aspect, not all checklist items can be achieved to the same degree at once. We, therefore, rely on the researchers to prioritize the checklist items and hence make trade-offs according to their research goals.
6.4 Step 2. Evaluation of the checklist

In the second step of our work, we conducted an evaluation of the checklist in a research practice context, i.e., with researchers that published survey research papers. This evaluation intended to assess the appropriateness of the checklist, i.e., how well it addresses the needs of the research community to assess survey-based research. In particular, we were interested in evaluating the completeness, relevance and fairness of the checklist. To address these goals, we formulated three evaluation questions:

EQ1. **Completeness:** Is the checklist missing any important aspect?

EQ2. **Relevance:** Does the checklist contain items that are not relevant for SE research? and

EQ3. **Fairness:** Is the assessment using the checklist too lenient or stringent?

6.4.1 Method

The goal of our evaluation was to verify whether researchers agreed with our independent assessment of their work, and furthermore whether they thought the checklist was complete and fair to assess survey research reports. To address such goal, we employed a mixed quantitative-qualitative approach, using our assessment through the checklist as the subject of study. We used a survey-based strategy to select and recruit participants for our evaluation. Similar strategies for evaluating methods and tools in SE context are described in [581].

A set of conditions favor the decision to adopt such an evaluation approach within the research community, such as:

- The subject of study is widely available, i.e., survey-based articles published in ESE journals and conferences. We can directly measure the compliance to the survey practices by assessing these objects with the support of our checklist.

- Researchers experienced with conducting surveys (i.e., corresponding authors of the above-mentioned papers) are likely to be interested in the checklist’s potential use. Their expert judgment is also essential to identifying limitations and to gather improvement suggestions.

- Some of the benefits from using the checklist to assess research (i.e., rigour and fairness) are difficult to quantify. Thus, open-ended questions are more likely to provide a deep understanding of the opinions and reasoning of the experts regarding such aspects.
An Empirically Evaluated Checklist for Surveys in Software Engineering

Our research evaluation process is based on a series of steps derived from a literature search [587] and survey-based recruitment and data collection [580]. Later, we analyzed the data according to both quantitative and qualitative synthesis procedures [563].

Selection and recruitment

Search strategy. At first, we identified survey-based articles that can be assessed using our checklist. We searched for potential candidates in nine venues (four journals and five conferences) publishing empirical research studies in SE, namely:

- **TSE** Transactions on Software Engineering
- **IST** Information and Software Technology
- **ESEJ** Empirical Software Engineering Journal
- **JSERD** Journal of Software Engineering Research and Development
- **ICSE** International Conference on Software Engineering
- **SEAA** Euromicro Conference on Software Engineering and Advanced Applications
- **IWSM-Mensura** International Workshop on Software Measurement
- **EASE** International Conference on Evaluation and Assessment in Software Engineering
- **ESEM** International Symposium on Empirical Software Engineering and Measurement

Despite well-established guidelines [580,586], our checklist also incorporate practices mentioned in recent guidelines (e.g. [569,569,590]). Thus, we opted for candidate papers that were published in the 5 most recent years (i.e., from 2012 to 2017), as they are more likely to incorporate such practices. From this database, we identified 3429 potential publications matching these characteristics.

Selection process. We further filtered the papers that mentioned the term “survey” in the title or abstract, thus narrowing the original list down to 177 candidates. We gathered these papers and selected them according to an inclusion criterion: *Does the paper clearly report survey-based research?* This resulted in 62 included papers.

Recruitment. Later, we invited by e-mail the corresponding authors of the selected papers to participate in our evaluation. Two of the corresponding
authors have more than one paper in our candidate list, thus we sent 60 invitations related to 62 resulting papers. The invitation letter presented the goal and the context of the research and also described the assessment procedure (see evaluation procedure, below).

Responses. Three invitation e-mails could not be received with the given e-mail address. Out of the 57 authors who received an e-mail, 22 agreed to participate. One of them consented in assessing two of the papers we asked for and also provided an extra paper which was not part of our dataset. As an incentive to recruiting the participant, we added the additional paper to our list.

Evaluation process

The process to evaluate our checklist consisted of:

1. collecting the referred paper and applying the checklist to assessing it;
2. providing the corresponding authors with the filled out checklist so that they could review our assessment; and finally
3. asking the corresponding authors to provide feedback regarding the checklist instrument and the resulting assessment.

The first author of this work carried out these three steps, later sharing the resulting assessment with the other authors. The resulting scores from our assessment using the checklist (see item 1, above) were aggregated into a dataset. This dataset was further analyzed in order to explore how many of the papers addressed each checklist item. Identifying patterns such as checklist items poorly addressed by most of the papers is essential for the next steps of our study. After receiving the participant’s feedback, we compared their review to the patterns we identified.

Out of the 22 corresponding authors who agreed to participate, 12 provided us with a feedback, which consisted of:

- a **review of our assessment**, in which the corresponding author can point out disagreements with our assessment, and and refine the assessment scores;
- **response to three opinion questions** regarding the checklist, as follows:

  1. Do you consider the checklist complete? If not, what should be included?
2. Is there anything you would like to remove, or do you think it is irrelevant?
3. Do you think our assessment by means of this checklist is fair? That is, was our assessment of the paper too rigid or too lenient?

Data analysis

Our analysis considered the feedback provided by the participants in the form of (1) a review of our assessment, and (2) answers to a set of opinion questions.

In order to analyse the review of our assessment, we gathered the notes and comments provided by the participants regarding each of the checklist items. These notes were used to assess the completeness and relevance (RQ1 and RQ2, respectively) of our checklist. In particular, we look for suggestions to improve the checklist, whether by removing, adding, or rephrasing. We responded to each comment and highlighted any action we took to improve the checklist based on the participants’ feedback (see Appendix 6.8.4).

Furthermore, we assessed the fairness of our assessment (RQ3) by computing the inter-rater agreement between the scores in our assessment and the ones reviewed by the corresponding authors. The inter-rater agreement is expressed in accordance with Cohen’s kappa coefficient.

We also aggregated the participants’ answers to three opinion questions into a common list\(^2\). These open-ended answers comprise the respondents own phrasing and reasoning regarding the three topics of our evaluation (i.e. completeness, relevance, and fairness).

We read each of the answers and assigned a value in a scale of yes/no/partial, representing their agreement with the question. We used both information types (i.e., assigned value and open-ended text) to answer our evaluation questions. Ultimately, we compared the participants’ opinions with the findings from the review of our evaluation in order to identify recurrent themes.

6.4.2 Threats to validity

Construct validity. A major threat to validity concerns the ability to assess the constructs with qualitative questions. We asked the participants to provide their own opinions regarding the checklist and our assessment. In particular, one of the participants questioned whether completeness could be assessed based on opinions.

\(^2\)Available at https://goo.gl/XE7uQF.
Conventionally, open questions are associated to subjective responses, which is likely to constrain the analysis of data. Besides the participants’ opinions, we support our findings with the scores resulting from our assessment. These scores are used to identify the practices often not reported or addressed.

**Interpretative validity.** Another potential threat to validity relates to the interpretation of the findings. In particular, we formulated three open-ended questions to collect participants’ opinions. The questions themselves are not bias-free, as they are formulated to extract a positive/negative response. As an example, “Do you consider the checklist complete?” received more positive than negative answers. To decrease this threat, the data analysis and interpretation of our evaluation study were conducted by the first author and discussed with the other co-authors.

**Reliability.** Our great involvement in constructing the checklist is likely to introduce personal biases on our assessment scores. We aimed to mitigate these by building a traceable chain of evidence. First, we assessed the selected papers and recorded notes to support the given scores. We later asked the corresponding authors to review our scores and notes, and to refine any disagreement they identified. We further computed the inter-rater agreement between ours and the participants’ scores, resulting in a very strong agreement (k = 0.91, according to weighted Cohen’s Kappa [559]).

The inter-rater analysis is commonly employed to assess the agreement between two raters independently. Here, our resulting assessment was shared with the corresponding authors prior to collecting their scores. Specifically, the participants were invited to review the scores they consider unfair and provide us refined ones. We assume that two reviewers using the checklist independently are not likely to achieve such stronger agreement. However, the results from the opinion questions also showed that the corresponding authors judged the assessment as mostly fair (see Section 6.4.1).

**External validity.** Our selection process aimed to identify a diverse set of survey-based articles, i.e. surveys in different areas and/or surveys of different quality. The sample of papers collected covers a wide range of SE topics, e.g., testing, modeling, and industry practice. These papers were peer-reviewed, so we assume they present a rigorous and sound description from the survey process. This assumption is supported by the results of our assessment, in which the average level of compliance is 65% of the items in our checklist. Thus, our sample is not diverse with regard to the methodological quality of the papers. Besides that, the participation of experienced researchers supports the generalization of our findings by expertise.
6.4.3 Results

Our assessment using the checklist

We applied our checklist to assess 24 papers reporting survey research. Each of the checklist items was ranked as fully addressed (F), partially addressed (P), not addressed (N), or not applicable (NA). A summary of our assessment is presented in Table A1.2.

Table A1.2: Summary of the combined scores obtained by the papers in our sample. Each row represents a checklist item, and the relative amount of papers (out of 24) ranked as fully addressed (F), partially addressed (P), not addressed (N), or not applicable (NA). The last column computes a compliance score based on how many papers address the related item.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>N</th>
<th>P</th>
<th>F</th>
<th>NA</th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Research Objectives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A</td>
<td>Are the research question(s)...</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>95.8%</td>
</tr>
<tr>
<td>1B</td>
<td>Is the research context defined?...</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>95.8%</td>
</tr>
<tr>
<td>1C</td>
<td>Are the needs for the survey...</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>95.8%</td>
</tr>
<tr>
<td>2. Study plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td>Is the survey process supported by guidelines?...</td>
<td>13</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>41.7%</td>
</tr>
<tr>
<td>2B</td>
<td>Is there a reflection on the need to update the research plan?...</td>
<td>19</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>20.8%</td>
</tr>
<tr>
<td>2C</td>
<td>Are the roles and responsibilities...</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>12.5%</td>
</tr>
<tr>
<td>3. Identify population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3A</td>
<td>Is the population characterized...?</td>
<td>13</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>45.8%</td>
</tr>
<tr>
<td>3B</td>
<td>Is the size of the population...</td>
<td>19</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>18.7%</td>
</tr>
<tr>
<td>4. Sampling plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>Is the kind of sample...defined?</td>
<td>8</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>56.2%</td>
</tr>
<tr>
<td>4B</td>
<td>Is the sample size calculated...</td>
<td>7</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>68.7%</td>
</tr>
<tr>
<td>4C</td>
<td>Are the sources of sampling...</td>
<td>1</td>
<td>2</td>
<td>21</td>
<td>0</td>
<td>91.7%</td>
</tr>
<tr>
<td>4D</td>
<td>Are the strategies and criteria to select units...</td>
<td>10</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>58.3%</td>
</tr>
<tr>
<td>5. Instrument design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5A</td>
<td>Is the type of instrument...defined?</td>
<td>1</td>
<td>1</td>
<td>22</td>
<td>0</td>
<td>93.7%</td>
</tr>
<tr>
<td>5B</td>
<td>Is the instrument design process...</td>
<td>5</td>
<td>2</td>
<td>17</td>
<td>0</td>
<td>75%</td>
</tr>
<tr>
<td>5C</td>
<td>Are the demographic questions...</td>
<td>2</td>
<td>2</td>
<td>20</td>
<td>0</td>
<td>87.5%</td>
</tr>
<tr>
<td>5D</td>
<td>Does particular care is taken to make the questions understandable...?</td>
<td>10</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>54.2%</td>
</tr>
</tbody>
</table>
### 6.4 Step 2. Evaluation of the checklist

#### Table A1.2 – Continued from previous page

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5E</td>
<td>Is the number and order of the questions taken in consideration?</td>
<td>16</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>5F</td>
<td>Is there a reflection on the type of responses...for the questions?</td>
<td>4</td>
<td>1</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>5G</td>
<td>If employing close-ended questions, are the standardized response...</td>
<td>1</td>
<td>3</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>5H</td>
<td>Is there a reflection on the adoption of additional sources...</td>
<td>18</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

### 6. Instrument validation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>6A</td>
<td>Is the validation process of the survey instrument detailed?...</td>
<td>6</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>6B</td>
<td>Is the instrument measuring what is intended?...</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>6C</td>
<td>In case of an electronic or online questionnaire, is the usability...</td>
<td>21</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6D</td>
<td>Are the results of the instrument validation discussed?...</td>
<td>10</td>
<td>1</td>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

### 7. Participant recruitment

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>7A</td>
<td>Are the strategies to select participants...</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>7B</td>
<td>Are the ancillary documents...</td>
<td>13</td>
<td>4</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>7C</td>
<td>If rewards or incentives to respondents are provided...</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>22</td>
</tr>
</tbody>
</table>

### 8. Response management

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>8A</td>
<td>Are the responses monitored?...</td>
<td>4</td>
<td>2</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>8B</td>
<td>Is there any action to be taken in case of non-responses...?</td>
<td>16</td>
<td>0</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

### 9. Data analysis

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>9A</td>
<td>Is the data validated...</td>
<td>16</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>9B</td>
<td>Is the method for data analysis...</td>
<td>2</td>
<td>2</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>9C</td>
<td>If statistical analysis is employed, is the hypothesis testing process...</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>9D</td>
<td>If using qualitative synthesis...</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>9E</td>
<td>If a stratified sample is defined...</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>21</td>
</tr>
</tbody>
</table>

### 10. Reporting

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>10A</td>
<td>Are the instrument and ancillary documents accessible...</td>
<td>5</td>
<td>1</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>10B</td>
<td>Has a discussion of both positive and negative findings...</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>10C</td>
<td>Are limitations of the study (e.g. threats to validity) discussed?</td>
<td>0</td>
<td>3</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>10D</td>
<td>Are the conclusions justified...</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.2</td>
</tr>
</tbody>
</table>
For each assessed item, we also added notes for possible improvements in the study’s documentation, e.g., due to missing information. As an example, in relation to the checklist item 2A, which assess the detailed procedures when designing a survey, we provided the following note to one of the participants: “The paper cited guidelines to survey research to characterize the sample and recruitment. It is not clear if the method provided in the guidelines are followed throughout all the research process.” In order to preserve the anonymity, we do not report the complete notes here. They were however shared with the corresponding authors.

The last column of the table presents a compliance score, i.e., the relative amount of papers that addresses the related checklist item. A score of 100% means that all papers were rated “F”. The NA ratings are not computed, and each P counts as half of a full score.

The average level of compliance with the checklist of selected papers in our sample is 65%. Some of the checklist items and groups presented better compliance, such as the items related to the research objectives (1A to 1C) and three out of four items related to reporting (10B to 10D). One expects any research work, regardless of research method employed, to meet these requirements.

The compliance score should be interpreted merely as an account of possible improvements to be taken into consideration for research practice. As an example, the characterization of the target population (checklist item 3A) is seldom reported, thus suggesting a need to foster the adoption of such practice recommended by the existing methodological guidelines.

Three checklist items are fully addressed by all the papers assessed. They cover practices such as incentives to responses (7C), qualitative synthesis (9D) and stratified data analysis (9E). These items are optional, and thus the assessment is rated not applicable (NA) for all the papers that do not employ such strategies. These results imply that researchers applying such strategies are likely to report them explicitly.

Among the checklist groups that are more scarcely addressed are:

2) study plan;
3) identify the population;
6) instrument validation; and
8) response management.

The low compliance scores show that the same kind of information is missing in several assessed studies. This implies that some of the recommended practices proposed by the guidelines are not followed. If we consider that this sample of
papers is a good representation of the overall survey-based research in SE, the low-compliance items point out to gaps that should be part of wider discussion so to see if they are relevant in survey research.

We later compared the participants’ scores to ours via inter-rater agreement. The resulting weighted Cohen’s Kappa coefficient $k = 0.91$ [559], suggesting a very strong level of agreement. Although we did not employ an independent rating process, the results from the inter-rater agreement are reinforced by the corresponding authors’ answers the opinion questions (see Section 6.4.3). Our assessment was considered fair by nine out of 12 the participants, and partially fair by the other three.

**Evaluation by the corresponding authors**

After assessing all the papers, we provided the corresponding authors with the filled out checklist and our notes. We then asked them to provide feedback based on our assessment. Out of the 22 corresponding authors contacted, 12 replied to our request, providing feedback regarding the completeness of the checklist, irrelevant checklist items, and fairness of our assessment.

We addressed the participants’ comments individually, responding to each issue in need of due attention and detailing the actions we took to improve the checklist. A subset of our responses are provided in Appendix 6.8.4, and the complete set is available online in [https://goo.gl/YDj1XA](https://goo.gl/YDj1XA).

**EQ1. Completeness:** Most of the participants (7 out of 12) agreed that the checklist was complete and included all the main aspects of survey-based research. Two participants thought that the checklist was partially complete, and it could be improved by clarifying a few items.

One participant highlighted their confidence that our method of creating the checklist was grounded in methodological publications, such as [586]. This information was not provided beforehand, so we assumed that the participant is familiar with such work, thus relating our checklist items to the recommended practice described in Kitchenham’s guidelines [586].

The three remaining participants who did not agree with the checklist completeness, raised issues such as: i) internal and external validity are not completely addressed in relation to the sampling plan and the instrument validation; ii) more details are needed for novice researchers using the checklist; and iii) validating completeness is not possible as an opinion. These aspects are addressed individually in our feedback document (see Appendix 6.8.4), as mentioned above.
EQ2. Relevance: Three participants mentioned irrelevant checklist items they believed should be removed:

2C) the checklist item addressing research roles and responsibilities was considered irrelevant for the report, but it could be part of the research plan (2B);

6A/6C) these two items should be combined, as they both address the instrument validation;

5H) using additional sources for data collection is optional, therefore if not mentioned in the paper it should be rated NA; and

7B) to provide ancillary documents (e.g., cover letter, invitation letter) is irrelevant to the research report.

The only issue raised by more than one participant is related to unifying 6A and 6C. The results of our assessment point out that most of our sample studies are in compliance with 6A (75%), but just a few (8.3%) actually address item 6C. We think that it is important to keep these two aspects separated, thus making explicit the needs for validating the usability of the questionnaire (see e.g., recommended practices P27, P28, P35, P39, P40, P41, P45). All the issues abovementioned are discussed in our feedback document (see Appendix 6.8.4).

EQ3. Fairness: Most participants (9 out of 12) considered our assessment being fair. Two of the participants also mentioned that despite rigid, the assessment was fair. Another one highlighted the need for instruments that promote rigorous assessment of the research methods. None of the participants described our assessment as completely unfair, although three of them pointed out that items we missed in our assessment were limitations to fairness.

We noted that two participants mentioned the lack of information due to size limitations of the publication. This issue is further highlighted in the comments of other participants (see Appendix 6.8.4). We sympathize with the participants’ concern regarding a fair assessment due to the size limitation. However, we stress the importance to provide all the details needed to properly assess the research based on its report. As a recommended practice, researchers are encouraged to make additional information (e.g., research diary, questionnaire instrument, ancillary documents) accessible to the target audience.
6.5 Discussion

6.5.1 Checklist Usage

In order to assess survey-based research, reviewers can employ the proposed checklist. Prior to assessment, we suggest verifying the availability of research process information (i.e., research report, survey instrument, and ancillary documents). Thereafter, each checklist item should be carefully read and then evaluated with respect to whether the question can be answered and was reflected on in the research report.

Several checklist items comprise two or more nested questions. Those items are intertwined and should not be assessed separately. Moreover, the checklist items can be addressed as partial coverage, due to the higher level of abstraction where answer is likely to be subjective. In such cases, we rely on the reviewers’ best judgment regarding the adoption of partial scores (i.e., 0.5).

It is possible to derive a scoring measure based on the checklist marks (e.g., 23 out of 38). However, we do not encourage the simple aggregation of scores in such a way, as it is likely to lead to a loss of assessment information. We suggest reviewers report the reasoning to score each question, thus highlighting the strengths and weaknesses of the assessed survey.

6.5.2 Implications to Research

The objective of our checklist is twofold: first, to audit reported survey-based research; and second, for supporting researchers in making research design decisions and reporting them. Ideally, both the researchers employing the checklist to plan and report their studies and the reviewers assessing the same research should obtain similar scores.

One can derive a tailored checklist instrument focus on the the particular needs. As an example, reviewers willing to use the checklist to audit reported survey studies could find our checklist too extensive. Moreover, external reviewers are potentially more interested to assess the evidence provided and quality of report. Those aspects are more closely related to the rationale R9) result-targeted and also to the validity aspects of R1) conclusion validity, R3) external validity, and R5) interpretative validity.

By filtering the checklist items according to this set of rationale, we obtain a tailored checklist as illustrated in Table A1.3. Note that the checklist is more strongly aligned to the process stages S9 Data analysis and S10 Reporting. These stages are more likely to be reported, as well as specific practices such
as 1A, 4A, and 4B. We further combined checklist items that relate to the decision-points we identified in Section 6.3.3, as this key practices are likely reported together. The resulting tailored checklist comprised 16 items.

Alternatively, this reflexive checklist can be used to improve the survey process; researchers are encouraged to think and reflect upon the questions they are aiming to use. In particular, trade-offs have to be made. The completeness of the survey as well as the ability to obtain a large and representative sample are desired, but also costly. Thus, as highlighted in the survey guidelines, the research process decisions have to be reflected upon with respect to cost-effectiveness. This is not to say that researchers should aim at minimizing the cost, but rather reflect on what is needed to fulfil the research goals. Researchers using our checklist are strongly encouraged to report the resulting scores along with their reflections about the checklist itself. We also foster independent evaluations to verify the appropriateness of the checklist to assess survey-based research by the research community.

Finally, the proposed checklist is intended to assess survey-based research in SE, but it has the potential to address different domains’ studies (e.g., social sciences). It is important to identify the differences of the survey-based process employed in SE and in other fields, thus evaluating the checklist in a cross-domain study.

Research Practice

During the checklist evaluation, we assessed 24 papers reporting survey research. The results of our assessment (Section 6.4.3) point to a list of recommended practices (see Section 6.3.3) that are scarcely addressed. We believe that by communicating these insights to the community, we can encourage researchers to consider the recommended practices in their research. The scarcely addressed practices are:

2. **Study plan:** A research plan or log book (see recommend practice P6) is important to guide the research efforts. This protocol should detail the responsibilities of each stakeholder (P52) and a timetable (P5). The document should be updated as new information becomes known, and ultimately make it accessible by the end of the research.

3. **Identify the population:** Very often the demographics of the participants are described, but scarce information is provided regarding the target population. An audience analysis (P7) is likely to identify and supply
6.5 Discussion

Table A1.3: An exemplary tailored checklist focused on reviewing aspects of evidence and reporting. These aspects are better represented by the checklist items related to rationale R1, R3, R5, and R9.

<table>
<thead>
<tr>
<th>Research Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A Are the research objective expressed in measurable terms? E.g. as research questions, or using the goal-question-metric approach.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identify the population</th>
</tr>
</thead>
<tbody>
<tr>
<td>3A Is the population or the survey’s target audience characterized (e.g. through audience analysis)?</td>
</tr>
<tr>
<td>3B Is the size of the population stated? If it is not possible to gather this data, are statistic estimates of the population drawn?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sampling plan and participant recruitment</th>
</tr>
</thead>
<tbody>
<tr>
<td>4A Is the kind of sample (i.e. probabilistic, non-probabilistic) defined? Obs. impact for data analysis, its representativeness and/or generalization should be discussed.</td>
</tr>
<tr>
<td>4B Is the sampling process described, and the resulting sample size presented?</td>
</tr>
<tr>
<td>4C Are the sources of sampling (e.g. particular databases or directories, open or restricted) defined? E.g. through a search plan.</td>
</tr>
<tr>
<td>4D/7A Are the strategies to select participants stated and implemented? E.g., through a sampling frame, as well as invitations, authorization codes, self-recruitment, or snowballing.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Response management</th>
</tr>
</thead>
<tbody>
<tr>
<td>8A Are the responses monitored? E.g. response rate, non-responsiveness, and drop-out questions. In case of inadequate response rate, the reasons for non-responses and drop-out items were investigated?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>9A Is the data validated prior to analysis? E.g. through checking inconsistent, incomplete and missing values.</td>
</tr>
<tr>
<td>9B Is the method for data analysis specified? Are the steps of the analysis process described? Are they suitable for the response formats collected?</td>
</tr>
<tr>
<td>5G/9C If statistical analysis is employed, is the hypothesis testing process documented and the standardized responses (i.e. nominal, ordinal, interval or ratio) stated? Appropriate scales should be assigned according to the mapped variables.</td>
</tr>
<tr>
<td>5C/9E Are the demographic questions formulated according to the audience? If a stratified sample is defined, are the data analysed according to demographics? Are there meaningful comparisons drawn from them?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>10A Are the instrument and ancillary documents accessible (e.g. URL link, external reference, appendix) to readers? If not, are the reasons for that discussed and convincing? If data resulting from the survey were disclosure, were anonymity and confidentiality of data discussed?</td>
</tr>
<tr>
<td>10B Has a discussion of both positive and negative findings been demonstrated? Are the discussion addressing the research question(s) or hypothesis? Does the discussion take into consideration the generalization of the findings?</td>
</tr>
<tr>
<td>10C Are the results of the assessment checklist reported? Are limitations of the study (e.g. threats to validity) discussed?</td>
</tr>
<tr>
<td>10D Are the conclusions justified by the results? Furthermore, are the implications and potential use of the results discussed?</td>
</tr>
</tbody>
</table>
these characteristics, and the census of practitioners and institutions could provide estimates of the population size (P23).

5. Instrument design: When designing the data collection instrument, one should carefully consider the order (P34) and amount (P36) of questions. Additional sources of data (P69), such as work repositories can provide means to cross-check the results.

6. Instrument validation: In general, the reports stated that some kind of validation of the questionnaire (e.g., a pilot) was performed. When employing online surveys, we should ideally check the usability of the questionnaire instrument. A series of practices (see e.g. P27, P28, P35, P39, P40, P41, and P45) can be used as a guideline or checklist to this validation. Furthermore, it is also important to report the improvements made as resulting of the instrument validation (P61).

7. Participant recruitment: Besides the need for making the research plan available, it is suggested to provide access to the standardized communication with participants. These documents include the invitation and cover letters (P60) as well as follow-ups and thank you letters (P56).

8. Response management: Besides the response rates, it is valuable to report any strategy used to improve responses, such as reminders (P70) or searching for additional databases (P16). These strategies are likely to affect the sample size, thus we should ideally discuss the implications of them for the study validity.

9. Data analysis: Before the analysis, researchers are encouraged to validate the data (P78) and check for inconsistent, incomplete or missing information (P74). There are several strategies to deal with missing data, from discarding to the imputation based on statistical models. Ideally, the implications due to employing such approaches should be described.

### 6.5.3 Comparison with Related Work

We previously identified related studies providing checklists for assessing empirical research in SE (see Section 6.2.4). The existing checklists in SE target mostly experiments [578, 583] and case study research [607]. Our proposed checklist is intended to address the issues of survey-based research and its specific stages, e.g., sampling, instrument design, and recruitment.
Our resulting checklist can be comparable to Stavru’s set of criteria to assess the thoroughness of surveys [603]. Similar to their work, we also systematically derived our instrument from the literature. We detailed herein the empirical processes to construct our checklist.

Our checklist differs from Stavru’s criteria, as it covers a larger set of practices and provide instructions based on the literature. The recommended practices often taken into account further activities of the survey research process. As an example, in relation to the research objectives, our checklist item A1: “Are the research objective expressed in measurable terms? E.g. as research questions, or using the goal-question-metric approach.” In this particular case, the objectives should describe what is intended to be measured, thus favoring data analysis, as proposed by [558,580,586,590].

The items in our checklist are not weighted, as the relative importance of each practice depends on the survey process, and researchers are foster to reflect on the key decision points. Besides that, we evaluated our checklist with experienced researchers. The evaluation identifies room for improvement, and we provide an updated post-evaluation version in Appendix 6.8.3.

More generic checklists [571,606] share some similarities with our proposed checklist. Similar questions are mostly related to the research objectives and reporting stages. This is expected since some actions (e.g., formulate research questions, define the scope, discuss the results and limitations of research) are inherent to all empirical research.

We considered the question formulation of the generic checklists when phrasing our proposed instrument. Researchers who used those similar checklists are likely to recognize the overlaps. This could be beneficial, as they can employ identical reasoning when assessing the common items. Similar questions provide the opportunity to compare the checklists or the studies assessed through them.

6.6 Conclusions

In this paper, we described a process to derive an instrument to assess survey-based research and its further evaluation. The motivation for such is grounded in the increased usage of survey-based research in SE. In a previous study, we identified several guidelines supporting the survey processes, but no instrument provided a checklist to assess their quality.

A set of 12 methodological studies provided qualitative data that was collected and analyzed through thematic analysis. The resulting themes resulted in sets of practices and rationales for carrying them out. We built the proposed
An Empirically Evaluated Checklist for Surveys in Software Engineering

checklist based on those extracted themes, reporting it through 38 questions organized by the survey process stages.

Later, we employed an empirical evaluation approach to collect experts' opinions of our checklist. We provided the experts with an assessment produced by applying the checklist to their own reported surveys. Overall, our instrument was evaluated as complete and the assessment as rigorous and fair. Issues regarding understandability and subjectivity of the checklist items were collected and based on this feedback, we update our proposed checklist.

We believe that the empirical software engineering community can benefit from our checklist for survey research. It can be a valuable asset for both researchers conducting and reporting survey-based studies, and for reviewers auditing survey reports.

As future work, we plan to investigate the potential benefits of using of the checklist by independent reviewers. We also intend to compare our checklist with other assessment tools (e.g. [578,583,607]) with respect to quality standards for empirical research.

6.7 References


[568] Rafael Maiani de Mello, Pedro Corrêa Da Silva, and Guilherme Horta Travassos. Investigating probabilistic sampling approaches for large-scale


An Empirically Evaluated Checklist for Surveys in Software Engineering


6.7 References


### 6.8 Appendix

#### 6.8.1 Co-occurrence tables

Table A1.4: Co-occurrence matrix of recommended practices (rows) according to the process stages (columns). Each cell contains a normalized coefficient, i.e., the highest co-occurrence value in each row is assigned a value 100 whereas the lowest value is 0. Cell shading illustrates the strength of the normalized co-occurrence coefficient, ranging from white (0) to teal (100).

<table>
<thead>
<tr>
<th>Recommended practices</th>
<th>Process stages</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Avoid too many objectives</td>
<td>100</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>P2</td>
<td>Goal-question-metric (GQM)</td>
<td>100</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>P3</td>
<td>Identify related studies</td>
<td>93</td>
<td>57</td>
<td>36</td>
<td>100</td>
<td>64</td>
<td>21</td>
<td>29</td>
<td>21</td>
<td>71</td>
</tr>
<tr>
<td>P4</td>
<td>Systematically follow procedures / guidelines</td>
<td>80</td>
<td>20</td>
<td>20</td>
<td>100</td>
<td>67</td>
<td>33</td>
<td>20</td>
<td>33</td>
<td>60</td>
</tr>
<tr>
<td>P5</td>
<td>Produce a schedule or timetable</td>
<td>100</td>
<td>20</td>
<td>20</td>
<td>54</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>P6</td>
<td>Keep a diary or log book</td>
<td>38</td>
<td>23</td>
<td>8</td>
<td>31</td>
<td>15</td>
<td>50</td>
<td>23</td>
<td>35</td>
<td>100</td>
</tr>
<tr>
<td>P7</td>
<td>Audience analysis</td>
<td>11</td>
<td>11</td>
<td>100</td>
<td>44</td>
<td>22</td>
<td>11</td>
<td>6</td>
<td>33</td>
<td>72</td>
</tr>
<tr>
<td>P8</td>
<td>Kind of sample: probabilistic</td>
<td>7</td>
<td>10</td>
<td>100</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>P9</td>
<td>Restricted databases, directories, groups, and subjects</td>
<td>63</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>P10</td>
<td>Source of sampling (SoS)</td>
<td>58</td>
<td>100</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>P11</td>
<td>Calculate sample size</td>
<td>8</td>
<td>10</td>
<td>100</td>
<td>2</td>
<td>11</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>P12</td>
<td>Update sampling frame based on additional collection of the same unit</td>
<td>38</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>P13</td>
<td>Snowballing selection</td>
<td>25</td>
<td>100</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>P14</td>
<td>Remove duplicates / redundant units</td>
<td>15</td>
<td>10</td>
<td>100</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>P15</td>
<td>Criteria for selecting the units of observation</td>
<td>10</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>P16</td>
<td>Open databases, social networks, digital libraries</td>
<td>16</td>
<td>38</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>P17</td>
<td>Search plan</td>
<td>63</td>
<td>13</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>P18</td>
<td>Randomize sample</td>
<td>9</td>
<td>30</td>
<td>100</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>P19</td>
<td>Increase sample size (or large-samples)</td>
<td>12</td>
<td>100</td>
<td>24</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>P20</td>
<td>Kind of sample: non-probabilistic</td>
<td>4</td>
<td>13</td>
<td>100</td>
<td>4</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Process stages</th>
<th>Recommended practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>P21 Define units of observation/analysis and search unit</td>
<td>16 16 23 100 26 10 13</td>
</tr>
<tr>
<td>P22 Create subgroups of the population (strata)</td>
<td>4 11 100 11 25 2 16 5</td>
</tr>
<tr>
<td>P23 Estimate the population size (through statistics)</td>
<td>18 100</td>
</tr>
<tr>
<td>P24 Recruitment approach: personalized selection</td>
<td>14 100 43 21 86 46 7</td>
</tr>
<tr>
<td>P25 Identify reasons for non-responses</td>
<td>63 100 50 44 63 88 13</td>
</tr>
<tr>
<td>P26 Recruitment approach: self-selection</td>
<td>20 100 33 53 20 53</td>
</tr>
<tr>
<td>P27 Completion measure or progress indicator</td>
<td>100</td>
</tr>
<tr>
<td>P28 Non-mandatory answers</td>
<td>100</td>
</tr>
<tr>
<td>P29 Prototype survey instrument</td>
<td>100</td>
</tr>
<tr>
<td>P30 Self-develop survey instrument</td>
<td>100</td>
</tr>
<tr>
<td>P31 Reuse survey instrument</td>
<td>40</td>
</tr>
<tr>
<td>P32 Create simple, unambiguous, actual and targeted questions</td>
<td>1 3 1 100 19 6 1</td>
</tr>
<tr>
<td>P33 Avoid intrusive and unethical questions</td>
<td>20 100 13</td>
</tr>
<tr>
<td>P34 Prioritize questions</td>
<td>100 15</td>
</tr>
<tr>
<td>P35 Questionnaire navigation</td>
<td>100 15</td>
</tr>
<tr>
<td>P36 Prune questions</td>
<td>39 100 33 11</td>
</tr>
<tr>
<td>P37 Define terminology</td>
<td>30 7 7 100 33 13</td>
</tr>
<tr>
<td>P38 Commercial survey instrument</td>
<td>25 100 50 17</td>
</tr>
<tr>
<td>P39 Questionnaire structure and format</td>
<td>6 100 37 9 17</td>
</tr>
<tr>
<td>P40 Survey instructions</td>
<td>9 100 63 13 19 6</td>
</tr>
<tr>
<td>P41 Allow resume answering</td>
<td>100 100 50</td>
</tr>
<tr>
<td>P42 Investigate demographic information</td>
<td>29 82 100 35 6 24</td>
</tr>
<tr>
<td>P43 Type of instrument self-administered/printed questionnaires</td>
<td>100 17 42 17</td>
</tr>
<tr>
<td>P44 Map items to research objectives</td>
<td>97 18 24 11 100 3 29 8</td>
</tr>
<tr>
<td>P45 Additional technological support</td>
<td>9 61 100 26 22 35 22</td>
</tr>
<tr>
<td>P46 Response format: open questions (unrestrictive)</td>
<td>7 100 4 48</td>
</tr>
<tr>
<td>P47 Error checking</td>
<td>100 67 22 56</td>
</tr>
</tbody>
</table>
### Table A1.4 – Continued from previous page

<table>
<thead>
<tr>
<th>Process stages</th>
<th>Recommended practices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td><strong>P48</strong></td>
<td></td>
</tr>
<tr>
<td>Improve response rates</td>
<td>28</td>
</tr>
<tr>
<td><strong>P49</strong></td>
<td></td>
</tr>
<tr>
<td>Response format: close-ended questions (standardize responses)</td>
<td>3</td>
</tr>
<tr>
<td><strong>P50</strong></td>
<td></td>
</tr>
<tr>
<td>Type of instrument: interviewer-administered</td>
<td>100</td>
</tr>
<tr>
<td><strong>P51</strong></td>
<td></td>
</tr>
<tr>
<td>Online questionnaires</td>
<td>7</td>
</tr>
<tr>
<td><strong>P52</strong></td>
<td></td>
</tr>
<tr>
<td>Define roles and responsibilities</td>
<td>85</td>
</tr>
<tr>
<td><strong>P53</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: compare instrument to a “gold standard”</td>
<td>100</td>
</tr>
<tr>
<td><strong>P54</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: inter-rater agreement measure</td>
<td>100</td>
</tr>
<tr>
<td><strong>P55</strong></td>
<td></td>
</tr>
<tr>
<td>Alternative versions of the instrument</td>
<td>11</td>
</tr>
<tr>
<td><strong>P56</strong></td>
<td></td>
</tr>
<tr>
<td>Ancillary document: thank you or follow-up letter</td>
<td>100</td>
</tr>
<tr>
<td><strong>P57</strong></td>
<td></td>
</tr>
<tr>
<td>Avoid leading participants</td>
<td>100</td>
</tr>
<tr>
<td><strong>P58</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: non-expert reviews</td>
<td>100</td>
</tr>
<tr>
<td><strong>P59</strong></td>
<td></td>
</tr>
<tr>
<td>Assessment through empirical research</td>
<td>10</td>
</tr>
<tr>
<td><strong>P60</strong></td>
<td></td>
</tr>
<tr>
<td>Ancillary document: cover letter</td>
<td>63</td>
</tr>
<tr>
<td><strong>P61</strong></td>
<td></td>
</tr>
<tr>
<td>Amend or update the instrument</td>
<td>18</td>
</tr>
<tr>
<td><strong>P62</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: focus group, discussion, reasoning</td>
<td>25</td>
</tr>
<tr>
<td><strong>P63</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: expert review</td>
<td>15</td>
</tr>
<tr>
<td><strong>P64</strong></td>
<td></td>
</tr>
<tr>
<td>Monitor response rate</td>
<td>79</td>
</tr>
<tr>
<td><strong>P65</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: test-retest</td>
<td>100</td>
</tr>
<tr>
<td><strong>P66</strong></td>
<td></td>
</tr>
<tr>
<td>Validation approach: pilot or pre-test</td>
<td>9</td>
</tr>
<tr>
<td><strong>P67</strong></td>
<td></td>
</tr>
<tr>
<td>Provide rewards or incentives to respondents</td>
<td>77</td>
</tr>
<tr>
<td><strong>P68</strong></td>
<td></td>
</tr>
<tr>
<td>Identify reasons for drop-outs</td>
<td>80</td>
</tr>
<tr>
<td><strong>P69</strong></td>
<td></td>
</tr>
<tr>
<td>Additional sources of data collection</td>
<td>100</td>
</tr>
<tr>
<td><strong>P70</strong></td>
<td></td>
</tr>
<tr>
<td>Send reminders</td>
<td>82</td>
</tr>
<tr>
<td><strong>P71</strong></td>
<td></td>
</tr>
<tr>
<td>Monitor respondents in real-time</td>
<td>75</td>
</tr>
<tr>
<td><strong>P72</strong></td>
<td></td>
</tr>
<tr>
<td>Discriminant analysis</td>
<td>100</td>
</tr>
</tbody>
</table>

Continued on next page
### Table A1.4 – Continued from previous page

<table>
<thead>
<tr>
<th>Recommended practices</th>
<th>Process stages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>P73 Categorize qualitative data (i.e. coding)</td>
<td>29</td>
</tr>
<tr>
<td>P74 Handling inconsistent, incomplete or missing data</td>
<td>15</td>
</tr>
<tr>
<td>P75 Data triangulation</td>
<td>83</td>
</tr>
<tr>
<td>P76 Statistical analysis</td>
<td>7</td>
</tr>
<tr>
<td>P77 Data comparison</td>
<td>29</td>
</tr>
<tr>
<td>P78 Data validation</td>
<td>11</td>
</tr>
<tr>
<td>P79 More than one question measuring the same aspect</td>
<td>23</td>
</tr>
<tr>
<td>P80 Data representation: tables, graphs, charts and plots</td>
<td>7</td>
</tr>
<tr>
<td>P81 Interpret the resulting data</td>
<td>13</td>
</tr>
</tbody>
</table>
6.8.2 Survey assessment checklist (pre-evaluation)

1. Research objectives
   A. Are the research question(s) specified?
   B. Is the context of research defined? Does it consider reasonable set of objectives? I.e., too many objectives require that particular considerations to the size and complexity of the questionnaire instrument are discussed. [P1, R7]
   C. Are the needs for the survey motivated? E.g., grounded on background and related studies. [P3, R5]

2. Study plan
   A. Is the survey process supported by guidelines? Does the researcher describe how the guidelines have been implemented? [P4, R7]
   B. Is there a reflection on the need to update the research plan? E.g., through keeping a research diary or log book [P6, R3]
   C. Are the roles and responsibilities of researchers and other stakeholders defined? E.g., through creating a schedule or timetable. [P5, P52, R7]

3. Identity population
   A. Is the population clearly characterized (e.g., through audience analysis)? [P7, R8]
   B. Is the size of the population stated? If it is not possible to gather that data, are estimates of the population drawn? [P23, R1, R9]

4. Sampling plan
   A. Is the kind of sample (i.e., probabilistic, non-probabilistic) defined? [P8, P30, R7]
   B. Is the sample size calculated and presented? Are the actions needed to obtain the sample described? [P21, P29, R1, R9]
   C. Are the sources of sampling (e.g., particular databases or directories, open or restricted) defined? E.g., through a search plan. [P9, P14, P16, P17, P21, R1, R7]
   D. Are the strategies and criteria to select units (of observation, of analysis and search unit) stated? E.g., through a sampling frame. [P35, P43, P51, P52, R7, R9]

5. Instrument design
   A. Is the type of instrument (i.e., self or interview-administered) defined? [P43, P50, R7]
   B. Is the instrument design process (acquisition, development, prototyping, testing, revision, description, validation) described? [P21, P29, R5, R9]
   C. Are the demographic questions formulated in a way to assure that they are posed adequately? If the stratification of the sample is planned, are the demographic question to be asked before entering the participants? [P22, P40, R5, R9]
   D. Does particular care is taken during the questions understandable and easy for the participant to provide an unbiased answer? E.g., P23, P37, P57, P64, R6, R7
   E. Is the number and order of the questions taken in consideration? [P53, P61, R7]
   F. Is there a collection on the type of response (e.g., open-ended, close-ended, or both) required for the question?” [P46, P48, R6, R9]
   G. If employing close-ended questions, are the standardized response formats (i.e., nominal, ordinal, interval or ratio) stated? [P44, R5, R9]
   H. Is there a reflection on the adoption of additional source for data collection? E.g., through the participant’s profile or supporting literature? [P90, R7]

6. Instrument validation
   A. Is the validation process of the survey instrument detailed? E.g., through piloting, pre-test, focus, cue groups, experts, expert or non-expert reviews. [P23, P54, P63, P65, P67, P68, P69, P74, R4, R6]

6A. Is the instrument measuring what is intended? Are the questionnaire items aligned to the research question(s)? [P44, R2]

6B. Is the instrument measuring what is intended? Are the questionnaire items aligned to the research question(s)? [P44, R2]

6C. Is the instrument measuring what is intended? Are the questionnaire items aligned to the research question(s)? [P44, R2]

6D. Is the results of the instrument validation discussed? After main problems have been identified, were the instrument updated/amended according to the validation results? [P41, R4]

7. Participant recruitment
   A. Are the strategies to select participants (stage 4. Sampling plan) implemented? E.g., through invitations, authorization codes, self-recruitment, or other strategies. [P13, R5, R7]
   B. Are the ancillary documents (e.g. invitation, cover and thank you letter) provided? If they were not produced, are the reasons for that discussed and convincing? [P76, P90, R6]
   C. If rewards or incentives to respondents are provided, are the reasons and implications (e.g. ethical concerns, bias) discussed? [P83, R7]

8. Response management
   A. Are the responses monitored? E.g., response rates, non-respondents, and drop-out reasons discussed. [P25, P44, P58, P71, R4]
   B. Is the action taken to be taken in case of non-response strategies discussed? [P78, R6]

8A. Are the responses monitored? E.g., response rates, non-respondents, and drop-out reasons discussed. [P25, P44, P58, P71, R4]

8B. Is the action taken to be taken in case of non-response strategies discussed? [P78, R6]

9. Data analysis
   A. Are the data validated prior to analysis? E.g., through checking consistency, incomplete and missing values. [P74, P76, R1, R9]
   B. Is the method for data analysis specified? Are the steps of the analysis process described? Are they appropriate for the response formats collected? [P46, P83, R5, R9]
   C. Is statistical analysis employed, is the hypothesis testing process clearly documented? Are the standardised responses clearly presented? E.g., through tables, graphs, charts and plots. [P46, P49, P72, P75, P86, R6]
   D. If using qualitative techniques (e.g. content-analysis, thematic or content analysis), is it clear how the categories/themes were derived from the data? [P46, P73, R5]
   E. If a stratified sample is defined (see 5C), are the data analysed according to demographics? Are there meaningful comparisons drawn from these? [P22, P77, R2, R9]

9A. Are the data validated prior to analysis? E.g., through checking consistency, incomplete and missing values. [P74, P76, R1, R9]

9B. Is the method for data analysis specified? Are the steps of the analysis process described? Are they appropriate for the response formats collected? [P46, P83, R5, R9]

9C. Is statistical analysis employed, is the hypothesis testing process clearly documented? Are the standardised responses clearly presented? E.g., through tables, graphs, charts and plots. [P46, P49, P72, P75, P86, R6]

9D. If using qualitative techniques (e.g. content-analysis, thematic or content analysis), is it clear how the categories/themes were derived from the data? [P46, P73, R5]

9E. If a stratified sample is defined (see 5C), are the data analysed according to demographics? Are there meaningful comparisons drawn from these? [P22, P77, R2, R9]

10. Reporting
   A. Are the instrument and ancillary documents accessible (e.g. url link, external reference, appendix) to readers? If not, are the reasons for that discussed and convincing? If data resulting from the survey were disclosed, are anonymity and confidentiality of data discussed? [P56, P90, R4, R7]
   B. Has a discussion of both positive and negative findings been demonstrated? Are the discussion addressing the research question(s) or hypothesis? Does the discussion take in consideration the generalization of the findings? [P34, R5, R9]
   C. Are the results of the assessment checklist reviewed? Are limitations of the study (e.g. threats to validity) discussed? [R5]
   D. Are the conclusions justified by the results? Furthermore, are the implications and potential use of the results discussed? [R1]

Figure 6.4: Survey assessment checklist proposed. This pre-evaluated version is later improved and updated (see Appendix 6.8.3).
6.8.3 Survey assessment checklist (post-evaluation)

1. Research objectives
   1A Are the research objective expressed in measurable terms? E.g. as research questions, or using the goal-question-metric approach.
   1B Is the research context defined? Does it consider a reasonable set of objectives? Of course, too many objectives require that particular aspects relating to a questionnaire’s size and complexity be discussed.
   1C Is the need for a survey research motivated (i.e. grounded on background and related studies)?

2. Study plan
   2A Is the survey process conducted based upon detailed procedures? Ideally, the survey process should also be based upon methodological guidelines.
   2B Is there a reflection on the need to update the research plan? E.g. through keeping a research diary or log book.
   2C Are the roles and responsibilities of researchers and other stakeholders defined? This information can be detailed in the research plan.

3. Identify population
   3A Is the population or the survey’s target audience characterized (e.g. through audience analysis)?
   3B Is the size of the population stated? If it is not possible to gather this data, are statistic estimates of the population drawn?

4. Sampling plan
   4A Is the kind of sample (i.e. probabilistic, non-probabilistic) defined? Of course, impact for data analysis, its representativeness and/or generalization should be discussed.
   4B Is the sampling process described, and the resulting sample size presented?
   4C Are the sources of sampling (e.g. particular databases or directories, open or restricted) defined? E.g. through a search plan.
   4D Are the strategies and criteria to select units (e.g. random selection) stated? E.g. through a sampling frame.

5. Instrument design
   5A Is the type of instrument (i.e. self- or interviewer-administered) identified? Are the strategies for participant recruitment and managing responses discussed?
   5B Is the instrument design process (acquisition, development, prototyping, versioning, reuse) described in the report?
   5C Are the demographic questions formulated according to the audience? If a stratification of the sample is planned, are the demographics adequate to characterize the participants?
   5D Has special care been taken to make the questions understandable by the respondents? E.g. through using a terminology familiar to the target population, or by providing a thesaurus.
   5E Has special care been taken to avoid intrusive and unethical questions? E.g. each bias may include questions that lead the respondent to a particular answer, or to expose personal data or behaviors.
   5F Is the number and order of the questions taken in consideration? In case of a potential bias related to the order of questions is identified, different versions of the instrument can be distributed.
   5G Is there a reflection on the type of responses (i.e. open-ended, close-ended or a mix of both) required for the questionnaire? Ideally, it should be possible to assess the type of each question, but the report could present the overall reasoning for the choices and provide a way to access the instrument.
   5H If employing closed-ended questions, are the standardized response formats (i.e. nominal, ordinal, interval or ratio) stated? Appropriate scales should be attributed to the questions according to the mapped variables.
   5I Is there a reflection on the adoption of additional sources for data collection? E.g. through the participant’s profile or supporting literature. Such additional sources may provide means for characterizing strata of participants or for validating data through cross-verification and triangulation.

6. Instrument validation
   6A Is the validation process of the survey instrument detailed? E.g. through piloting, pre-test, focus groups, experiments, expert or non expert reviews.
   6B Is the instrument measuring what is intended? Are the questionnaire items mapped to the research question(s)?
   6C Is the instrument measurable? E.g. questionnaire navigation, instructions of use, option to resume answering, progress indicator, required/non-input, aesthetics, and layout.
   6D Are the results of the instrument validation discussed? After the main problems have been identified, were the instrument updated/amended according to the validation results?

7. Participant recruitment
   7A Are the strategies to select participants (stage 4. Sampling plan) implemented? E.g. through institutions, authentication codes, self-recruitment, or snowballing.
   7B Are the auxiliary documents (e.g. invitation, cover and thank you letters) proposed? If they were not provided, are the reasons for that discussed and convincing?
   7C Is there a reflection on the need to update the research plan? E.g. through keeping a research diary or log book.

8. Data analysis
   8A Are the responses monitored? E.g. response rate, non-responses, and drop-out questions. In case of inadequate response rate, the reasons for non-responses and drop-out items were investigated?
   8B Is there any action to be taken in case of non-responses (e.g. reminders)? If reminders are employed, is the process for selecting and inviting new participants described? Moreover, are the implications of reminders discussed? I.e. changes in the sample size are likely to impact the heterogeneity and generalizability of data.

9. Data analysis
   9A Is the data validated prior to analysis? E.g. through checking inconsistent, incomplete and missing values.
   9B Are the methods for data analysis or outcome(s) and the steps of the analysis process described? Are they suitable for the response format collected?
   9C Is statistical analysis performed, is the hypothesis testing process documented and the standardized responses presented? E.g. through tables, graphs, charts and plots.
   9D Is using qualitative synthesis (e.g. meta-ethnography, thematic or content analysis), is it clear how the categories/themes were derived from the data?
   9E If a stratified sample is defined, are the data analysed according to demographics? Are there meaningful comparisons drawn from them?

10. Reporting
    10A Are the instrument and auxiliary documents accessible (e.g. URL link, external reference, appendix) to readers? If not, are the reasons for that discussed and convincing? If data resulting from the survey are disclosed, were anonymity and confidentiality of data discussed?
    10B Has a discussion of both positive and negative findings been documented? Are the discussion addressing the research question(s) or hypothesis? Does the discussion take into consideration the generalization of the findings?
    10C Are the results of the assessment checklist reported? Are limitations of the study (e.g. threats to validity) discussed?
    10D Are the conclusions justified by the results? Furthermore, are the implications and potential use of the results discussed?

Figure 6.5: Survey assessment checklist after evaluation (see Section 6.4 from Chapter 6). A digital version of the checklist is available at https://tinyurl.com/se-survey-checklist.
6.8.4 Suggestions to improve the checklist (excerpt)

Here we present a sample of the feedback provided by the participants (i.e., corresponding authors) of our evaluation in the professional context (Section 6.4). The comments are listed according to the checklist item they are related to. For each comment, we present our responses and actions we took to address the mentioned topics. A complete list detailing all the comments is provided at https://goo.gl/jNXx7U.

1B) Two comments regarding the understandability of this checklist item:

i) What type of context and limitations should be described? Many questions would benefit from having a more detailed guide along with the checklist.; and

ii) The term "limitations of scope" may be misleading. Reading quickly I first thought that you referred to whether the study scope has some limitations to be able to answer RQs (related to study validity).

Response: We agree that term “limitations of scope” can leave room for interpretations. It could also be misleading, as limitations are often described as threats to the validity of a study.

Action: To improve the checklist understanding, we rephrased item 1B, as follows: “Is the research context defined? Does it consider a reasonable set of research objectives? Obs. too many objectives requires that particular aspects relating to a instrument’s questionnaire’s size and complexity be discussed.”.

2) Two comments regarding the description of this checklist item:

i) The sub-questions for me do not address the main question of whether a survey is appropriate. 2A-C are more about what is reported, rather than whether survey is the right method. That for me is more about whether other approaches were considered etc. Roles and responsibilities I would generally not note in a paper.; and

ii) (...) I am not convinced that the three questions that are included in this category would be enough to assess if a survey study research design is appropriate to address its research aims (as it is stated in the question). The fact that guidelines are followed, a research diary is kept and responsibilities are defined does not guarantee that the research design is appropriate to answer the research questions. (...)
**Response:** We agree with the authors that the description of checklist item 2 does not match what is assessed in sub-items 2A-2C. These sub-items assess whether the study plan is accessible and complete, instead of “appropriate”.

In order to assess whether the survey design is appropriate to address its research objectives, we designed a specific question (see 6B), that assess if the questionnaire items are related to the research questions described in the study plan.

The recommended practices for the checklist item 2 are: (i) to provide a survey plan document; (ii) to report the guidelines used; and (iii) to detail the responsibilities of each researcher. These aspects are important to allow for the study to be reviewed and replicated.

**Action:** We updated phase 2’s description to match what is assessed by items 2A-2C: “Study plan: Is the survey design accessible and complete?”.

**2C) Two comments regarding the relevance of this checklist item:**

i) *Is this information really relevant for the report? I think this information would fit better into a protocol or research plan than in the report itself*; and

ii) *It sounds irrelevant (e.g., how relevant it is for assessing the survey itself to know timetables)?*

**Response:** We agree with the authors that a schedule or timetable is not relevant for assessing the quality of the survey. However, these artifacts can provide means to assess the roles and responsibilities of researchers in the survey process [580]. As suggested by one of the participants, the information regarding the roles and responsibilities of each researcher could be provided in a survey plan document. We highlight here the need to make this document accessible to reviewers [?].

**Action:** We removed the references to schedule and timetables in the item 2C, instead stating of that “This information can be detailed in the research plan (see item 2B)”.

**5H) Three comments regarding the understandability of this checklist item:**

i) *I didn’t understand it.;*

ii) *It makes no sense for me. If man don’t mention another source of information it means there is not. Why do you assume that there is another unmentioned source?; and*
iii) I don’t fully follow this question. Do you mean data triangulation so findings from the survey are triangulated with other data?

Response: The authors pointed out a very important understanding issue. Our proposed checklist accounts for the need to discuss if additional sources are required, e.g., to characterize strata of the participants, or to cross validate the data related to the investigated phenomenon from multiple sources [580,590]. As an example, after the survey, the findings can be compared to other sources, such as personal profile information or related work.

Action: We added a note on 5H making explicit the reasons to adopt additional sources of data collection, as follows: “Such additional sources may provide means for characterizing strata of participants or for validating data through cross-verification and triangulation”.

10A) Is data disclosure / open data also a criterion? I think it should be as people should be pushed in the general direction of open science to foster reproducibility.

Response: The author suggests that, besides the ancillary documents we already mentioned in the checklist, the resulting data from the survey is also make available. We see the value on that, but acknowledge potential implications due to anonymity and confidentiality that should be taken into consideration. Therefore, we rely on the judgment of researchers conducting the survey to provide a discussion on such aspect.

Action: We added a note on 10A making explicit that the data disclosure can be provided and thus should be discussed. The new item is “If data resulting from the survey are disclosure, does a consideration about the anonymity and confidentiality of data is discussed?”.

10A) Two additional comments regarding applicability of this checklist item:

i) I can’t imagine including things like invitations, thank you notes, etc. My ethics approval process requires so much documentation, I would need a separate 10 pages just for all of the information I provide to participants. ; and

ii) not all materials can be added to an appendix especially with paper length limitation.

Response: We agree with the author that limitations due the publication size are likely to constrain the amount of information provided in a paper.
As an alternative, the additional documents mentioned in 10A could compose a “survey research package”, available as a web reference provided in the paper [557].
ABSTRACT

**Background.** Software Engineering (SE) research, like other applied disciplines, intends to provide trustful evidence to practice. To ensure trustful evidence, a rigorous research process based on sound research methodologies is required. Further, to be practically relevant, researchers rely on identifying original research problems that are of interest to industry; and the research must fulfill various quality standards that form the basis for the evaluation of the empirical research in SE. A dialogue and shared view of quality standards for research practice is still to be achieved within the research community.

**Objectives.** The main objective of this thesis is to foster dialogue and capture different views of SE researchers on method level (e.g., through the identification and reasoning on the importance of quality characteristics for experiments, surveys and case studies) as well as general quality standards for Empirical Software Engineering (ESE). Given the views of research quality, a second objective is to understand how to operationalize, i.e. build and validate instruments to assess research quality.

**Method.** The thesis makes use of a mixed method approach of both qualitative and quantitative nature. The research methods used were case studies, surveys, and focus groups. A range of data collection methods has been employed, such as literature review, questionnaires, and semi-structured workshops. To analyze the data, we utilized content and thematic analysis, descriptive and inferential statistics.

**Results.** We draw two distinct views of research quality. Through a top-down approach, we assessed and evolved a conceptual model of research quality within the ESE research community. Through a bottom-up approach, we built a checklist instrument for assess survey-based research grounded on supporting literature and evaluated ours and others’ checklists in research practice and research education contexts.

**Conclusion.** The quality standards we identified and operationalized support and extend the current understanding of research quality for SE research. This is a preliminary, but still vital, step towards a shared understanding and view of research quality for ESE research. Further steps are needed to gain a shared understanding of research quality within the community.