SEGRESS: Software Engineering Guidelines for REporting Secondary Studies

Barbara Kitchenham, *Member, IEEE,* and Lech Madeyski, *Senior Member, IEEE* and David Budgen, *Member, IEEE*

Abstract—*Context*: Several tertiary studies have criticized the reporting of software engineering secondary studies. *Objective*: Our objective is to identify guidelines for reporting software engineering (SE) secondary studies which would address problems observed in the reporting of software engineering systematic reviews (SRs). *Method*: We review the criticisms of SE secondary studies and identify the major areas of concern. We assess the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement as a possible solution to the need for SR reporting guidelines, based on its status as the reporting guideline recommended by the Cochrane Collaboration whose SR guidelines were a major input to the guidelines developed for SE. We report its advantages and limitations in the context of SE secondary studies. We also assess reporting guidelines for mapping studies and qualitative reviews, and compare their structure and content with that of PRISMA 2020. *Results*: Previous tertiary studies confirm that reports of secondary studies are of variable quality. However, *ad hoc* recommendations that amend reporting standards may result in unnecessary duplication of text. We confirm that the PRISMA 2020 statement addresses SE reporting problems, but is mainly oriented to quantitative reviews, mixed-methods reviews and meta-analyses. However, we show that the PRISMA 2020 item definitions can be extended to cover the information needed to report mapping studies and qualitative reviews. *Conclusions*: In this paper and its Supplementary Material, we present and illustrate an integrated set of guidelines called SEGRESS (Software Engineering Guidelines for REporting Secondary Studies), suitable for quantitative systematic reviews (building upon PRISMA 2020), mapping studies (PRISMA-ScR), and qualitative reviews (ENTREQ and RAMESES), that addresses reporting problems found in current SE SRs.

Index Terms—evidence-based software engineering, reporting guidelines, systematic reviews, quality reviews, mapping studies, mixed-methods reviews, threats to validity, risk of bias, quality assessment, PRISMA 2020

1 Introduction

The goal of this article is to introduce and justify the SEGRESS guidelines that we have developed for reporting secondary studies in software engineering (SE). The SEGRESS guidelines are based on the PRISMA 2020 standard, which was developed to support the reporting of medical and healthcare-related systematic reviews.

The main reason for developing SEGRESS was to address criticisms of SE systematic review reports raised in recent tertiary studies (see Budgen et al. [1], Zhou et al. [2], Ampatzoglou et al. [3], Yang et al. [4]). Criticisms include problems in finding the information required, such as recommendations [1], lack of standards for assessing the validity of secondary studies ([2] and [3]), and problems with study quality assessment [4]. In Section 2, we summarise the criticisms reported in these studies in more detail. This both justifies the need for SE reporting guidelines that are suitable for software engineering researchers, and also identifies essential information that such guidelines need to ensure is reported.

In Section 3, we introduce the PRISMA 2020 statement, which is the current international standard for reporting SRs in healthcare. Since the original SE guidelines for software engineering systematic reviews ([5], [6]) were based on healthcare guidelines, it seems plausible that PRISMA 2020 could be of use to SE researchers. In this section, we confirm that once the terminology used in PRISMA 2020 is explained, it addresses all the issues raised in Section 2.

However, the developers of PRISMA 2020 make it clear that the statement is intended for quantitative SRs, mixed-method reviews, and meta-analysis. This limitation on the scope of PRISMA 2020 is a major barrier to its adoption by researchers in software engineering, because secondary studies in SE are often mapping studies (i.e., secondary studies that aim to classify the literature related to a research topic) or qualitative reviews (i.e., secondary studies that use qualitative methods to synthesize primary study results) [7]. To overcome this limitation, we have developed SEGRESS as an extension of PRISMA 2020 that incorporates guidelines for mapping studies and qualitative reviews.

In Section 4, we show that mapping studies can be reported using PRISMA 2020 by omitting some of the standard items related to risk of bias (i.e., quality assessment and certainty assessment), and providing extended explanations to confirm that the synthesis for mapping studies is restricted to simple charts and category counts.

It is much more difficult to provide standards for reporting qualitative reviews than for mapping studies, not least because there is no definitive standard for *conducting* qualitative systematic reviews. However, in Section 5, we introduce two guidelines that have been developed to report

- *B. Kitchenham is with the School of Computing and Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK.ORCID: 0000-0002-6134-8460*
- *Lech Madeyski is with the Department of Applied Informatics, Wroclaw University of Science and Technology, Wyb.Wyspianskiego 27, 50370 Wroclaw, Poland.*
  E-mail: Lech.Madeyski@pwr.edu.pl ORCID: 0000-0003-3907-3357
- *David Budgen is with the Department of Computer Science, Durham University, Durham DH1 3LE, UK. ORCID: 0000-0001-7143-0241*

Manuscript received Month XX, 2022. (Corresponding author: Lech Madeyski)

qualitative systematic reviews. We show that the specific items in these reporting guidelines relate to the PRISMA 2020 items, and confirm that the preliminary guidelines for reporting qualitative studies can be integrated with PRISMA 2020.

In Section 6, we present the SEGRESS (Software Engineering Guidelines for REporting Secondary Studies) guidelines, which are based on the PRISMA 2020 statement and have definitions of the items that have been extended to make them suitable for mapping studies and qualitative reviews. To support the use of the guidelines and to provide some initial validation of their value, we provide Supplementary Material [8] that includes extended explanations of each item and examples of how the items can be reported taken from existing quantitative and qualitative software engineering secondary studies. The material also includes a hypothetical running example (based on revising an existing SE SR) to illustrate how related information reported in different parts of a SR report need to be organized. We discuss our results and present our recommendations for reporting SE secondary studies in Section 7.

The main research questions and research goals addressed in this article and the procedures used to develop SEGRESS are reported in different sections of this article and its Supplementary Material [8]. The main issues addressed by each section are shown in Table 1. Readers who are interested primarily in SEGRESS may find it useful to read Section 3 and Section 6 before looking at Section 2, Section 4 and Section 5, which explain the development of SEGRESS. It is important to read Section 3 before looking at the SEGRESS guidelines in order to understand the terminology used in PRISMA 2020 which we have adopted in SEGRESS. We also strongly recommend consulting the Supplementary Material [8] when trying to use the SEGRESS guidelines. The additional explanations and examples are critical to understanding how to apply the guidelines.

2 Tertiary Studies Criticizing SR Reporting Practices

In this section, we discuss four recent SE tertiary studies that discussed SR reporting practices in software engineering.

2.1 General Reporting Issues

Budgen et al. [12] undertook a tertiary study to investigate the extent to which secondary studies in software engineering provided results that could be used for industrial or educational purposes. To address this issue, they selected and studied 178 SE systematic reviews published between 2010 and 2015. As a result of studying those reviews, they identified a set of 12 "lessons learned" that were aimed at identifying good practice for reporting SE systematic reviews [1].

They reviewed the SRs included in their tertiary study in the context of the DARE[1] (Database of Attributes of Reviews of Effects) assessment criteria:

1) Are the review's inclusion and exclusion criteria described and appropriate?
2) Is the literature search likely to have covered all relevant studies?
3) Did the reviewers assess the quality/validity of the included studies?
4) Were basic data/studies adequately described?
5) Were the included studies synthesised?

Budgen et al. organized their lessons against the five DARE criteria together with one other criterion related to their goal of identifying any conclusions which could be used as guidelines for education or industry practice. Based on the 12 lessons, they specify nine items they consider to be essential information to specify in an SR report. These items are shown in Table 2, which specifies the SR issue of concern in column 1 and the information that should be reported in column 2. We also map the issues to items in the PRISMA 2020 statement in column 3, which we discuss in Section 3.

Budgen et al.'s study makes it clear both that there are problems with reporting the conduct and results of many SR processes, and that it would be useful to have a well-defined structure for a secondary study report so that authors know what information to provide in each section, and readers know where to look for particular pieces of information.

2.2 Reporting Threats to Validity

Both Zhou et al. [2] and Ampatzoglou et al. [3] criticized the reporting of threats to validity in SE secondary studies.

Zhou et al. [2] investigated the threats to validity reported in systematic reviews published from 2004 to mid-2015. They found 316 secondary studies of which they identified 178 as SRs, 132 as systematic mapping studies, and six as meta-analyses. From each paper (including mapping studies), they extracted the reported threats to validity and classified them against the standard threats to validity for empirical studies (construct, internal, external, conclusion, content, concurrent, predictive, statistical). They also identified the stated impact of each threat and the mitigation action adopted. They concluded that while most SRs reported internal validity and reliability issues, few reported construct validity and external validity issues. Additionally, they reported that methods for addressing threats to validity were seldom reported.

Ampatzoglou et al. [3] performed a study very similar to [2]. They identified 449 secondary studies from the time period 2007-2016, including both systematic reviews and mapping studies. They found 165 papers reporting threats to validity. From both the data they collected and the consideration of the SE systematic review process guidelines, they constructed a checklist of 22 threats to validity in systematic reviews grouped into three major categories: Study Selection, Data Validity, Research Validity. For each threat, they identified one or more mitigation actions. They used a panel of experts to assess the relative effectiveness of the different mitigation actions. They recommend their checklist to authors of secondary studies both to identify potential threats and to report threats to validity. They also suggest that readers, including reviewers, should use the checklist to assess the validity of secondary studies.

Ampatzoglou et al. [3] make a very good point that classic threats to validity, as described in the social sciences (see, for example, [13]), are not generally applicable to systematic reviews. However, their approach has some limitations (which are shared by Zhou et al.'s study), in particular:

1. https://www.crd.york.ac.uk/CRDWeb/

TABLE 1
SEGRESS Development

Research Question or Goal	Research Approach	Outcome
Do we need SR reporting standards?	Review critiques of SE SRs and identify critical reporting issues.	Yes, we need reporting standards. See Section 2 and the summary of results in Table 2.
Is PRISMA 2020 (the medical and health care standard) a possible solution?	We reviewed the scope of PRISMA 2020 and confirmed that it addresses problems reported in Section 2.	Yes PRISMA 2020 is a possible standard, see Section 3.1, *but* PRISMA 2020 is of limited value for SE because of SR type restrictions. In addition, it requires understanding of terminology related to Risk of Bias and Quality of Evidence, see Section 3.3.
Can PRISMA 2020 be adapted for Mapping Studies?	All three authors independently assessed whether PRISMA for scoping reports (PRISMA-ScR) could be used for mapping studies and could be mapped to PRISMA 2020.	We agreed it was possible. See Section 4.1 for our methodology and Section 4.2 and Table 5 for the main results.
Can PRISMA 2020 be adapted to support Qualitative Reviews?	All three authors independently assessed whether two proposed standards for reporting qualitative reviews (ENTREQ [9] and RAMESES [10]) could be mapped to PRISMA 2020.	We agreed it was possible, but the special requirements of qualitative reviews need to be understood, see Section 5.1. See Section 5.2.1, Table 7 and Table 8 for our assessment methodology and results.
Develop the SEGRESS checklist	Update PRISMA 2020 items with descriptions that apply to Mapping Studies and Qualitative Studies based on Table 5, Table 7 and Table 8	SEGRESS development is reported in Section 6.1, and the SEGRESS checklist is reported in Table 9. A preliminary validation of SEGRESS is reported in Section 6.2
Provide additional practical support for SEGRESS users.	Provide extended explanations suitable for for SE researchers based on PRISMA 2020 [11] and standards for qualitative reviews ([10] and [9]) together with SE related examples.	See the Supplementary Material [8].

TABLE 2
Essential Information For Systematic Review Reports [1]

Review aspect	Information Required	PRISMA item (Section 3)
Inclusion/ Exclusion	The rules for both inclusion and exclusion should be clearly stated.	5
Inclusion/ Exclusion	How the rules were applied, and any difference between reviewers were resolved should be described.	8
Inclusion/ Exclusion	The number of papers remaining at each stage of selection should be reported.	16a
Searching	All of the search mechanisms used should be clearly reported.	6, 7
Searching	The period covered by the search should be explicitly stated, and the dates when any searches were performed should be reported.	5, 6
Quality Assessment	When performed, the intended use as well as the checklist items should be reported.	13e, 13f, 15
Quality Assessment	How quality assessment was undertaken, and how any differences between reviewers were resolved needs to be explained.	11, 15
Synthesis	Where performed, the form of the synthesis adopted for specific research questions should be described, and the reasons for its use should be given.	13a
Outcomes	Key findings should be clearly reported, together with any information related to the "strength of evidence" that applies to them.	22, 23b, 23d

1) The information the researchers extracted answered the question "What threats to validity *are* reported", it does not answer the question "What threats to validity *should be* reported".
2) As noted by Ampatzoglou et al., systematic review guidelines were explicitly designed to mitigate many threats to validity in secondary studies, for example, requiring extensive searches to avoid publication bias, and having multiple reviewers independently address tasks such as searching, selection, data collection and quality assessment to avoid researcher bias. Furthermore, the SR methodology should be reported in the Methods section, which should also specify and justify any planned deviations from SR process guidelines. Thus, Ampatzoglou et al. may have under-estimated the extent to which SE researchers have actually avoided many threats to validity.
3) There was no explicit consideration of the difference between a systematic review and a mapping study, although many threats to validity may be different. For example, among Data Validity threats, Ampatzoglou et al. include *The selection of classification system is biased* which is a mapping study issue, and among Research Validity threats they include *Lack of comparable studies*, that is a quantitative systematic review issue that is irrelevant for mapping studies which do not investigate the outcomes of empirical studies.

The practical problem of adopting Ampatzoglou et al.'s recommendations is that a checklist is supposed to represent a complete set of items that must all be addressed. However, attempts to apply the complete checklist in an isolated Threats to Validity section can lead to duplicate reporting of issues that may have already been covered in the Methods and Results sections. This can be seen in Ampatzoglou et al.'s paper by comparing their Methodology Section (Section 3) with the content of their Threats to Validity Section (Section 7). For example, they report both their search strings and the fact that they checked their set of studies against other tertiary studies in both their description of their tertiary study search process (Section 3.2) and in their discussion of

threats to validity (Section 7.1).

Our review of [2] and [3], provides support for Budgen et al.'s call for a well-defined structure for SR reports. In addition, it seems necessary to have:

- A *theoretical rationale* for what we should and should not report in a threats to validity section. For example, it should include only issues that have not been fully explained in other parts of the paper.
- An approach to reporting validity threats for systematic reviews and mapping studies that properly reflects their differences and similarities. For example, since mapping studies do not synthesize the outcomes of primary studies, there can be no threats to validity associated with statistical meta-analysis such as primary study heterogeneity, publication bias, or generalizability.

2.3 Quality Assessment

Yang et al. [4] performed a tertiary study to assess the use of quality assessment in SE SRs which updated the results of a previous study [14] and covered the period 2004-2013. Their review included 241 SRs published between 2004 and 2018 that used a quality assessment instrument. They report that the use of qualitative assessment had improved since their first tertiary study. In particular, they conclude that "the aims of quality assessment are more concise, the instruments used are more diverse and rigorous and the criteria more thoughtful".

Yang et al. investigated why researchers assessed primary study quality and found that 46 primary studies did not explain why they had collected quality assessment data. They identified and classified 202[2] reasons identified by the authors of the remaining 195 SRs for using quality assessment as being related to:

- Selection: to provide more extensive inclusion and exclusion criteria. This was the most frequently identified reason (i.e., 54% of the identified reasons).
- Interpretation: to guide the interpretation of the findings and determine the strength of inference, which accounted for 16% of reasons
- Investigation: to understand the current state of research, which accounted for 14% of reasons.
- Validation: to ensure that only studies of good quality are included, which accounted for 10% of reasons.
- Weighting: as a means of weighting the importance of individual studies when results are being synthesized, which accounted for 5% of the reasons.

Yang et al.'s study confirms that there is still some confusion about the reasons for undertaking quality assessment. They also make two further important points about quality assessment which we will refer to again later in this paper:

1) The QA instruments for tertiary studies and secondary studies are different. Since the primary studies in a tertiary study are secondary studies, tertiary studies can all use the DARE criteria as quality assessment criteria. For standard systematic reviews, the quality assessment criteria need to reflect the specific methodology or methodologies used in the primary studies, which can cause problems if different primary studies in the same SR have used many different research methods
2) It is important to consider whether or not to conduct quality assessment.

2. The individual counts reported by Yang et al. make it clear that some SR authors had more than one reason for performing a quality assessment.

2.4 Conclusions for Reporting SE Systematic Reviews

All of the tertiary studies discussed in this section have reported justifiable criticisms of current reporting practices for secondary studies used by software engineering researchers. However, *ad hoc* and uncoordinated changes in reporting practices aiming to change individual aspects of SR reporting may cause confusion about what is to be reported in other related sections, and might also cause SE systematic review terminology to deviate from existing standards. In our opinion, we need standards for reporting the different forms of systematic review that provide an overall structure for systematic review reports and that define and explain the issues that need to be reported.

In the remainder of this paper, we discuss whether the PRISMA 2020 statement for reporting systematic reviews can provide the basis for software engineering reporting guidelines. The guidelines used in software engineering for *performing* systematic reviews arose from the medical and healthcare guidelines, so we assumed that PRISMA 2020, which was designed for reporting medical and health care systematic reviews [15], was likely to be a useful starting point for SE reporting standards.

3 The PRISMA Statement

Given the problems with reporting SRs that have been raised by many SE researchers, we believe that it is important to adopt a standard for reporting SRs. In this section, we review the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement and assess how far it could provide a means of addressing these issues.

PRISMA 2020 is an update to the original PRISMA statement [16] which was published in 2009. The original PRISMA statement was widely adopted and led to the development of a series of related standards, as shown in Table 3. The PRISMA statement for scoping reviews, called PRISMA-ScR [17], is discussed in Section 4 where we discuss whether PRISMA 2020 can be used for reporting mapping studies. Other PRISMA extensions include guidelines for the development of a SR protocol, called PRISMA-P [18], guidelines for the format of a SR study abstract, called PRISMA-A [19], and guidelines for the SR search process, called PRISMA-S [21]. In a separate initiative, the Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group developed standards for assessing the strength of recommendations from systematic reviews ([22] and [23]).

PRISMA 2020 is reported, in full, in two documents [33] and [34]. Page et al. [33] undertook a mapping study to identify reporting guidelines for SRs published after 2009. They identified 60 sources containing 221 unique items. They finally identified 175 items that could be used as a

TABLE 3
PRISMA (Preferred Reporting Items for Systematic Reviews) and Related Standards

ID	Name	Scope	Derivation
PRISMA [16]	Preferred Reporting Items for Systematic Reviews	Reporting quantitative systematic reviews and meta-analysis	
PRISMA-ScR [17]	PRISMA Extension for Scoping Reviews	Reporting scoping reviews	Based on PRISMA after removing items related to synthesis and risk of bias
PRISMA-P [18]	Preferred reporting items for systematic review and meta-analysis protocols	Developing protocols for systematic reviews and meta-analyses that will be reported using PRISMA	Based on specifying all the items in a PRISMA–compliant SR
PRISMA-A [19]	PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts	Specifying abstracts for PRISMA-based systematic reviews in journals and conferences	Based on PRISMA. Updated in PRISMA 2020 [20]
PRISMA-S [21]	PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews	Supports three items in PRISMA: Information sources, Search Strategies and Study Selection Results (items 7, 8 and 14 in PRISMA [16] and items 6, 7, and 16 in PRISMA 2020 [15])	Based on PRISMA and an extensive expert-opinion based development process, including consultation with developers of PRISMA 2020.
GRADE [22], [23], [24]	Grading of Recommendations Assessment, Development, and Evaluation	Assessing the strength of recommendations made in systematic reviews, health technology assessments and clinical practice guidelines. Supports PRISMA 2020 item 22 Certainty of Evidence.	The original ideas were presented in [22] and have been revised and refined in a series of articles produced by a technical working group, see ([25], [26], [27] , [28], [29], [30], [31], [32]).
PRISMA 2020 [20]	The PRISMA 2020 statement: an updated guideline for reporting systematic reviews	Reporting quantitative systematic reviews, evaluation studies, meta-analysis, and mixed methods. Includes an update to PRISMA-A	Based on PRISMA after substantial research ([33] and [34])

comprehensive item bank for future reporting guidelines. They used this information to identify potential revisions to the original PRISMA statement [34]. They invited 220 systematic review methodologists and journal editors to complete a survey on suggested modifications and received 110 replies. The recommendations from the survey were reviewed by a 21-person group that produced a draft revision to the original PRISMA statement, which was then further refined based on feedback from co-authors and a convenience sample of 15 systematic reviewers.

We considered using PRISMA 2020 as a possible standard for SE SRs because it is important to adopt reporting guidelines that are based on a widely adopted international standard that has been subject to a rigorous development process. Another advantage of adopting PRISMA 2020 is that SE researchers can keep their terminology aligned with other empirical disciplines.

PRISMA 2020 identifies 27 *items*, some of which contain sub-items, where each item and sub-item identifies an issue that should be reported. Items 1-23 map broadly to the standard content of a scientific paper: Title, Abstract, Introduction, Methods, Results, and Discussion. The final four items relate to ethical scientific practice, specifically: protocol registration, funding, reporting competing interests, and data availability. The basic PRISMA 2020 checklist is reported in [15] and is explained in more detail in [20] with examples taken from existing systematic reviews.

PRISMA 2020 was designed primarily for "systematic reviews of studies that evaluate the effects of health interventions, irrespective of the design of the included studies". However, the authors confirm that the checklist items:

- are applicable to non-health-related interventions,
- are suitable for mixed-methods reviews, i.e., reviews that include quantitative and qualitative studies, although other reporting guidelines should be consulted for qualitative synthesis, e.g., [9].

In addition, the guidelines for constructing abstracts (PRISMA-A [19]) was revised by Page et al. [20] to make its wording consistent with PRISMA 2020.

The basic scope and definitions of the PRISMA 2020 items are shown in Table 4. From this table, we can see that once it is appreciated that the term *risk of bias* is a replacement for the term *quality assessment*, and that *limitations* is a replacement for *threats to validity*, PRISMA 2020 recommends reporting all aspects of the SR process and the results of applying that process. In particular, the basic ordering of the items requires users to specify the methods they used to perform all the required SR processes in the Methods section, the results of applying those methods in the Results section, and to discuss the findings of the review and any recommendations in the Discussion section. Thus, it appears to address all of the reporting problems raised by SE researchers.

3.1 Preliminary Assessment of PRISMA 2020

As an initial feasibility check of the potential value of PRISMA 2020, the three authors explicitly investigated whether PRISMA 2020 addressed the issues raised by Budgen et al. [1].

3.1.1 Preliminary Assessment Method

Kitchenham circulated a document including the 9 items shown in Table 2. We independently identified any PRISMA 2020 item that addressed each issue with any explanation necessary to support our assessment. Kitchenham collated the individual assessments and circulated a spreadsheet that identified each of the individual assessments and any related comments. All authors reviewed their assessments and those of the others individually, and made any changes they felt were necessary, adding any relevant comments to support their assessments. The revised assessments were again collated by Kitchenham and circulated. We then discussed (by e-mail) any remaining disagreements.

3.1.2 Preliminary Assessment Results

The results of assessing whether PRISMA 2020 addresses the reporting problems raised by Budgen et al., are shown in the final column of Table 2 where we identify the PRISMA items

where required information should be reported. During our discussions, our main disagreements concerned whether or not:

- PRISMA addressed all the issues related to quality assessment. Disagreements arose because some of the specific issues raised by Budgen et al. [1] were only discussed in the more detailed explanations and examples of PRISMA items.
- Budgen et al. [1] recommend the dates of both the start and end of any searching process be explicitly defined. In fact, by default, PRISMA assumes that there is no lower bound on the search and that the end date of the search will be reported in PRISMA item 7. If researchers have set other date-based limits on the search, these should be reported (and justified) under item 5 as eligibility criteria.

However, our discussions confirmed that PRISMA 2020 is not simple to understand and needs additional explanations and examples from software engineering to be suitable for software engineering researchers. In the following sections, we discuss some of the practical issues that need to be resolved if SE researchers are to be able to adopt the PRISMA 2020 statement. In particular, we discuss how PRISMA 2020 addresses threats to validity in Section 3.2 and quality assessment issues in Section 3.3. In addition, Section 3.4 discusses what is meant by mixed-methods in the context of systematic reviews, and Section 3.5 discusses some practical difficulties involved in reporting studies with multiple findings associated with different subsets of primary study outcomes.

3.2 Threats to Validity

PRISMA 2020 does not mention *Threats to Validity* and instead refers to limitations of the evidence included in the review (see item 23b) and limitations of the review process used (see item 23d).

Page et al. [20] explain the request to "Discuss any limitations of the evidence included in the review" (see item 23b) as follows:

> *"Discussing the completeness, applicability, and uncertainties in the evidence included in the review should help readers interpret the findings appropriately. For example, authors might acknowledge that they identified few eligible studies or studies with a small number of participants, leading to imprecise estimates; have concerns about risk of bias in studies or missing results; or identified studies that only partially or indirectly address the review question, leading to concerns about their relevance and applicability to particular patients, settings, or other target audiences. The assessments of certainty (or confidence) in the body of evidence (item 22) can support the discussion of such limitations."*

Thus, this item should report problems with the findings from the synthesis due to limitations arising from the scope or reliability of the primary studies. We discuss what is meant by *confidence in the body of evidence* and its relationship to quality assessment in more detail in Section 3.3.

Page et al. explain the request to "Discuss any limitations of the review process used" (see, item 23c) as follows:

> *"Discussing limitations, avoidable or unavoidable, in the review process should help readers understand the trustworthiness of the review findings. For example, authors might acknowledge the decision to restrict eligibility to studies in English only, search only a small number of databases, have only one reviewer screen records or collect data, or not contact study authors to clarify unclear information. They might also acknowledge that they were unable to access all potentially eligible study reports or to carry out some of the planned analyses because of insufficient data. While some limitations may affect the validity of the review findings, others may not."*

This explanation confirms that researchers should report and discuss any decisions they have made that are in conflict with the standard systematic review guidelines. An important issue is that researchers need to discuss the implications of any deviations from the standard SR guidelines in terms of their likely impact on the review findings.

3.3 Quality Assessment, Risk of Bias and Certainty Assessment

PRISMA 2020 does not make reference to *Quality Assessment*, instead it uses the term *Risk of Bias* (RoB). It considers both RoB associated with individual studies and RoB associated with syntheses. In addition, it also recommends the use of *Certainty Assessment*, in order to assess the confidence in (or quality of) the body of evidence relating to a specific finding.

The important difference between RoB and quality assessment for individual studies is that RoB is about identifying potential *methodological flaws* that can bias the outcome of primary studies, whereas quality is about whether the research was performed as well as possible. For example, in medical research, there is strong *empirical* evidence that failure to blind both participants and experimenters can bias experimental outcomes [35]. We have no reason to assume software engineering is exempt from such problems. For example: Ciolkowski [36], in the context of inspections; and Shepperd et al. [37], in the context of fault prediction models; both reported that the outcomes of their meta-analyses revealed evidence of experimenter bias. In the context of SE experiments, it is seldom possible to blind the participants and experimenters. Therefore, although SE researchers may perform other aspects of their experiments to the highest possible standard (i.e., the quality may be high), lack of blinding remains a significant RoB.

As shown in Figure 1, RoB assessments of primary studies are intended to be used for three purposes:

1) sensitivity analysis, where researchers assess the extent to which specific findings are dependent on the RoB of influential studies;
2) investigation of causes of heterogeneity, where researchers investigate whether high or low risk of bias is associated with positive or negative findings;
3) as part of the certainty assessment process, which is discussed below.

Risk of bias with respect to synthesis is mainly an issue of bias due to missing results from various forms of *non-reporting* bias, also known as publication bias. Publication bias occurs because negative results may not be submitted for publication (the so-called "file drawer problem"), or may be published in sources that make them more difficult to find. For example, they may not be accepted by high-status journals, may be published in national, non-English sources, or may not be published in a timely manner. Publication bias is an important element of certainty assessment.

TABLE 4
The PRISMA 2020 Statement

Section	PRISMA Item	Description
Title	1	Identify the report as a systematic review.
Abstract	2	See the PRISMA 2020 Abstract checklist
Introduction		
Rationale	3	Describe the rationale for the review in the context of existing knowledge.
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.
Methods		
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for synthesis.
Information sources	6	Specify all databases, registers, web sites, organizations, reference lists and other sources, to be searched or consulted. Specify the date when each source was last searched or consulted.
Search Strategy	7	Present full search strategies for all databases, registers and websites, including any filters and limits used.
Selection Process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and, if applicable, details of automation tools used in the process.
Data collection process	9	Specify the method used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and, if applicable, details of automation tools used in the process.
Data items	10a	List and define all outcomes for which data was sought. Specify whether all results that were compatible with each outcome in each study were sought (e.g., for all measures, time points, analyses), and, if not, the methods used to decide which results to collect.
	10b	List and define all other variables for which data was sought (e.g., participant and intervention characteristics, funding source). Describe any assumptions made about any missing or unclear information.
Study Risk Of Bias Assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and, if applicable, details of automation tools used in the process.
Effect Measures	12	Specify for each outcome the effect measure(s) (e.g., risk ratio, mean difference) used in the synthesis or presentation of results.
Synthesis Methods	13a	Describe the process used to decide which studies were eligible for each synthesis.
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling missing summary statistics, or data conversions.
	13c	Describe any methods used to tabulate or visually display results of individual studies and synthesis.
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of heterogeneity, and the software packages(s) used.
	13e	Describe any methods used to explore possible causes of heterogeneity among study results.
	13f	Describe any sensitivity analysis conducted to assess robustness of the synthesized results.
Reporting Bias Assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting bias).
Certainty Assessment	15	Describe methods used to assess certainty (or confidence) in the body of evidence for an outcome.
Results		
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.
	16b	Cite studies that met many but not all inclusion criteria ("near-misses") and explain why they were excluded.
Study characteristics	17	Cite each included study and present its characteristics.
Risk Of Bias In Studies	18	Present assessments of risk of bias for each included study.
Results of Individual Studies	19	For all outcomes, present for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g., confidence/credible interval), ideally using structured tables or plots.
Results Of Synthesis	20a	For each synthesis, briefly summarize the characteristics and risk of bias among contributing studies.
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g., confidence/credible interval) and measures of statistical heterogeneity, if comparing groups, describe the direction of the effect.
	20c	Present results of all investigations of possible causes of heterogeneity among study results.
	20d	Present results of all sensitivity analysis conducted to assess the robustness of synthesized results
Risk of Reporting Bias in Synthesis	21	Present assessments of bias due to missing results (arising from reporting biases) for each synthesis assessed.
Certainty of Evidence	22	Present assessment of certainty (or confidence) in the body of evidence for each item assessed.
Discussion		
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.
	23b	Discuss any limitations of the evidence included in the review
	23c	Discuss any limitations of the review process used.
	23d	Discuss implications of the results for practice, policy and future research.
Other information		
Registration and Protocol	24a	Provide registration information for the review, incl. register name & registration number, or state that the review was not registered.
	24b	Indicate where the review protocol can be accessed or state that the protocol was not prepared.
	24c	Describe and explain any amendments to information provided at registration or in the protocol.
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.
Competing Interests	26	Declare and competing interests of the review authors.
Availability Of Data, Code and Other Materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.

Certainty assessment refers to methods used to assess the *confidence in synthesized results*. The results of a specific synthesis depend on the *set of primary studies* used in the specific synthesis, and the aim of certainty assessment is to evaluate the level of confidence that can be placed in the results of a specific synthesis (referred to as assessing *the quality of evidence*) and any recommendations based on that synthesis (referred to as assessing the *strength of the recommendation*). Thus, the difference between certainty assessment and primary study RoB is due to the "unit" being evaluated. For study risk of bias the "unit" is the individual primary study, for certainty assessment, it is the specific synthesis obtained from the set of relevant primary studies. This means that it is possible for different findings from the same study to be assessed as having different quality of evidence. For example, in software engineering studies, some researchers may assess the impact of a technique only on development effort, whereas others may assess the impact on both development effort and task duration. If only the primary studies assessed as having high risk of bias are the ones that report task duration outcomes, any assessment of the quality of evidence associated with findings associated with duration will be lower than the quality of evidence associated with development effort findings.

Certainty assessment relies on the results obtained by assessing RoB for individual studies, as well as the risk of publication bias and other factors such as imprecision in effect size estimates (i.e., large effect size variances) and inconsistency (i.e., heterogeneity) among study results. The relationships between risk of bias and certainty assessment are shown in Figure 1 and are described in more detail in Section 3.3.3.

3.3.1 Risk of Bias for Primary Studies

The Cochrane Handbook [35] provides advice on assessing risk of bias for individual primary studies, although the most detailed information it provides is related to randomized controlled trials, which is of little relevance in the context of empirical software engineering (since we do not often undertake randomized formal experiments in industry settings). The risk of bias associated with non-randomized designs and quasi-experiments in Chapters 22-25 is more relevant and is based on four different risk domains:

1) *Confounding*, which occurs when a factor other than the intervention of interest could have caused the effect. For example, in a field trial of the effectiveness of detailed design inspections, the project manager restricted design inspections to complex components. This confounded the use of the intervention with the complexity of the component, which complicated the interpretation of the study results [38].
2) *Selection bias*, which occurs when some eligible participants, or some outcome events, are excluded in a way that leads to systematic bias in the outcomes. For example, in a study of the use of formal methods, the researchers restricted eligibility to the most capable students [39].
3) *Information bias*, which may be introduced if intervention status is wrongly classified, or if outcomes are wrongly classified or measured with error. For example, if software engineers are asked to adopt a new testing method which is more complex or time-consuming than their current method, they may revert to the current method in order to complete their assigned task. A case in point is a study of inspection methods where the authors mention that adherence to the perspective-based method was sketchy [40]. The issue of process conformance in the context of Test-Driven Development experiments was discussed, e.g., in [41] and [42].
4) *Non-reporting bias*, for example, experimenters reporting only outcomes that have significant results.

In addition, there are numerous well-known examples of poor practice that may introduce bias into quantitative studies, such as small sample size, over-simplistic tasks, lack of effect sizes and confidence intervals, and multiple statistical tests, whether as a result of many different outcome variables or testing many different subsets of the data (see, for example, [43], [44] and [45]).

In their tertiary study, Yang et al. [4] discussed the quality assessment instruments used in SE systematic review reports in the time period 2004-2018. They reported that the most commonly used criteria related to four areas:

- Rationality, which is related to the study rationale, its context and its research questions.
- Rigour, which is related to the choice of the research methodology and the way in which it was applied.
- Credibility, which is related to the clarity and validity of the reported results and the extent to which they are supported by the evidence, and the relationship between experimenters and participants.
- Contribution, which is related to the value of the findings both for industry and academia.

The items identified in the Cochrane Handbook are related to the issues identified by Yang et al., but they are less abstract, which means they may be easier for reviewers to understand and use in a RoB assessment.

3.3.2 Risk of Bias for Synthesized Outcomes

RoB for findings obtained by synthesizing results from a set of primary studies is basically the risk of publication bias. Although there are statistical methods for assessing the extent of possible publication bias in the context of meta-analysis, in other situations it is assessed in terms of the depth and breadth of the search process, and the appropriateness of eligibility criteria. Although RoB for synthesized outcomes is treated as a separate issue to Certainty Assessment in PRISMA 2020, we found considerable overlap between the concepts, which is also evident in the presentation of the concepts in the Cochrane Handbook. We discuss certainty assessment in more detail below.

3.3.3 Certainty Assessment

For Certainty Assessment, Page et al. [20] mention GRADE (Grading of Recommendations Assessment, Development and Evaluation) as an example instrument. GRADE has been adopted by the Cochrane Handbook for assessing certainty (or quality) for a *body* of evidence (see [35], chapter 14). In [46], Dybå and Dingsøyr discuss use of the original GRADE approach [22] in the context of software engineering. However, there has been no discussion of the impact of the latest revisions of the GRADE approach in the SE literature.

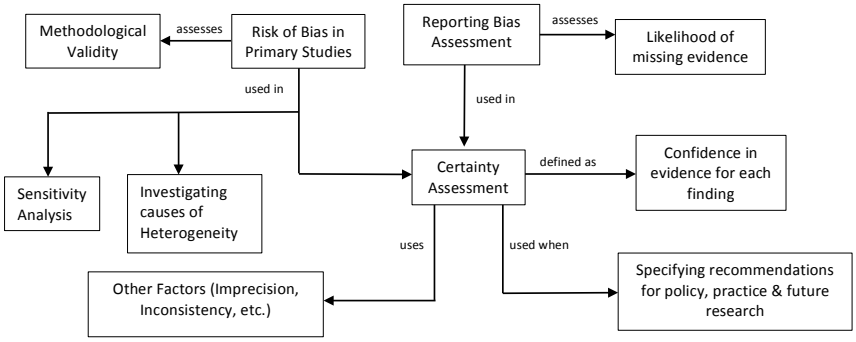

Fig. 1. Relationship between Risk of Bias and Certainty Assessment

The Cochrane Handbook (which was updated in 2019) and the series of articles by Guyatt et al. [24], published in the Journal of Clinical Epidemiology, describe the current version of GRADE, but use terminology that may be unfamiliar to SE researchers. In this section, we discuss the GRADE concepts and explain how they can be applied to quantitative software engineering SRs. We also discuss GRADE for qualitative studies in Section 5.1.3.

The recent GRADE revision specifies four levels of the certainty for a body of evidence related to a given finding: high, moderate, low and very low (see [24] and [26]). GRADE considers five domains, each of which must be assessed against each of the four levels:

1) *Risk of bias of individual studies*. The idea is to assess whether the primary studies associated with a particular finding are assessed as low risk of bias, leading to high certainty in the finding, or share some methodological bias(es) that increase risk of bias and reduce certainty in the finding. In SE, small-scale, student-based laboratory experiments should initially be regarded as having a high risk of bias, and thus provide very low certainty for evidence. In contrast, large experiments with industry practitioners (such as, [47], [48]) could be assessed as having relatively low risk of bias (the classification scheme proposed by Höst et al. [49]). Furthermore, field studies, whether qualitative or quantitative, should initially be regarded as high quality evidence but could be re-assessed to a higher risk of bias if their methodology was particularly weak. For more details about specific empirical methods, Felderer and Travassos [50], Easterbrook et al. [51] and Stol and Fitzgerald [52] all provide useful discussions of the range of empirical methods used in contemporary software engineering research and their strengths and weaknesses. For data-based studies, Menzies and Shepperd [53] have identified 12 "bad smells" as indicators of potential problems in software analytics papers. In addition, issues such as analyses based on extremely old or untrustworthy data sets (for example, the NASA dataset, see [54] and [55]), or use of unreliable metrics (such as mean magnitude relative error, see [56], [57], and [58]) form a risk of bias for the synthesis of fault prediction and cost estimation studies respectively.
2) *Publication bias*. In the context of meta-analysis, the risk of missing studies can be investigated analytically using techniques such as funnel plots that correct for publication bias. However, for SRs that cannot apply meta-analysis, the risk of publication bias can usually only be addressed by the stringency of the search process, but can also consider factors such as whether the primary studies are dominated by small positive studies [59]. This is discussed in more detail in the Supplementary Material [8].
3) *Imprecision* is related to the confidence intervals associated with overall effect size estimates. Confidence intervals that do not exclude the null hypothesis would usually lead to a reduction in the certainty associated with the body of evidence [29].
4) *Inconsistency*. This relates to whether the direction of the effect size is consistent across the individual studies [30]. The body of evidence is downgraded if studies give inconsistent results. In qualitative studies, strong disagreements between the findings of studies with similar contexts would be an indication of inconsistency. Agreement between qualitative and quantitative studies about whether an intervention is beneficial or not would also be an indicator of high certainty of evidence.
5) *Indirectness* relates to whether the studies directly test the concept of interest or are inferred from indirect comparison [31]. Causes of indirectness relevant in the context of SE SRs are: participants may not be representative (e.g., students without industrial experience may not respond to a new technique in the same way as practitioners); the reported intervention differs from the intervention of interest (e.g., we are interested in the impact of test-first development but only have studies that consider test-first in a maintenance context): or the outcome measures differ from those of primary interest (e.g., Turner et al. [60] point out that a great majority of the papers that evaluated the Technology Acceptance Model evaluated the model only against the user's *intention to use* not against actual measures of system usage).

The process of evaluating findings against the GRADE domains is subjective. It should usually be done by several reviewers independently. Each reviewer should provide an explanation for their assessment of each domain for each separate review finding. Differences among reviewer assessments should be discussed and, if necessary, mediated until agreement is reached.

3.4 PRISMA for Mixed-Methods Studies

Mixed methods reviews are methods for synthesizing and integrating qualitative evidence with intervention reviews [61]. Page et al. [15] explicitly state that PRISMA 2020 is relevant for mixed-methods systematic reviews. However, they also point out that PRISMA does not itself provide reporting guidelines addressing the presentation and synthesis of qualitative data and other guidelines need to be consulted. Specifically, they refer readers to [9] and [62].

Paraphrasing the discussion of Noyes et al. [61] on the use of qualitative evidence for complex health interventions, in the context of software engineering, the reasons for including qualitative evidence are as follows:

- better understanding of why and how an intervention works;
- identifying associations between the broader social and technological environment, within which software engineers work, and the interventions that are implemented;
- understanding of attitudes towards, and experiences of, interventions by the software engineers who are expected to adopt them; and
- increasing understanding of the software development process factors that are most likely to impact the success or failure of an intervention.

Many SE researchers have commented that large-scale industrial software engineering is a complex activity. For example, Curtis et al. [63] describe the context of industrial software engineering as a "layered behavioral model" involving individual programmers, the teams in which they work, the projects on which they work, the organization that employs them, and the business sector in which the organization does business. At the level of software development companies, this view is echoed both by the systems dynamics modeling method proposed by Abdel-Hamid et al., [64], and Belady and Lehman's laws of system evolution [65]. Furthermore, software engineering research methods are increasingly focused on evaluations performed in industrial settings (e.g., [66]). Thus, we would expect the mixed-methods approaches recommended in the Cochrane Handbook [61] to be of particular value when synthesizing results from such studies.

Harden et al. [67] provide a useful discussion of methods and tools that can be used to integrate qualitative and process evaluation evidence within intervention effectiveness reviews. They identify five approaches, but all start by tabulating the findings from quantitative studies with qualitative factors either reported in the quantitative studies or available from other qualitative reviews/studies. The simplest approach is then to perform a narrative review of the tabulated information to explore the heterogeneity between the quantitative study findings and look for research gaps. For example, in systematic reviews addressing two different aspects of cost estimation, Jørgensen [68] and Kitchenham et al. [69] both identified contextual factors that could be used to help cost estimators to decide which estimation method to use, given their specific circumstances. Furthermore, with sufficient quantitative studies reporting comparable outcome measures, the factors identified as influencing outcomes can be tested statistically. Such statistical tests can be performed one at a time (see [36]), or, with sufficient studies, can test multiple factors together (see [37]). Harden et al. [67] also mention:

- constructing or refining a logic model which is a graphical representation of process factors that influence the outcome of a complex intervention;
- developing or refining theories about how the intervention should be implemented;
- using what they refer to as "Qualitative comparative analysis" which aims to investigate multiple factors across many different contexts.

It is hard to find examples of SE researchers adopting these more sophisticated approaches to integrate qualitative models with quantitative review findings. Albeit, we are aware of a synthesis method called Structured Synthesis Method (SSM) [70], [71], [72] that is intended to allow the combination of different types of evidence.

3.5 Iteration and Repetition

In this section, we discuss practical issues that cause problems when using the PRISMA 2020 guidelines. In particular, there is a change in the unit of discussion, as the statement items change from discussion of primary study data and risk of bias to discussion of each review finding (i.e., each answer to a research question or outcome from a specific synthesis activity). This reporting problem is further complicated because assessment of the certainty of the evidence needs to be based on the set of primary studies that contribute to each of the specific findings, rather than being a single overall assessment of credibility of all the primary studies.

3.5.1 Iteration

One reporting problem with PRISMA 2020 is that items 18 and 19 appear to assume a linear order for reporting all primary study RoB data and outcome data. In contrast, all item 20 sub-items and item 21 require reporting results for each *finding*. Thus, there is some iteration among items, but it is not clearly defined; for example, PRISMA item 19 in [20] implies that all forest plots used in the systematic review are reported together, which can include different plots for different outcome variables and which may relate to different subsets of the primary studies. Then, for all the sub-items of item 20 and item 22, it asks authors to discuss the meta-analysis results and the relevant RoB assessment and certainty assessment, for each finding. In practice, it may be preferable to report the forest plots, textual explanation of the meta-analysis findings plus any sensitivity analysis or heterogeneity assessment, and the certainty assessment together for each separate synthesis.

For SE SRs, it is quite common to separate the primary studies into separate groups that address different aspects of a topic (see, for example, [73], [74]). In such cases, it may increase readability to iterate through items 17 to 22 for each subgroup. However, it may still be useful to precede any iterative group-based reporting with single tables for primary study data (item 17) and RoB (item 18) for data that are collected for all primary studies irrespective of subgroup.

3.5.2 Repetition

Another issue related to readability is how to organize reporting of related issues without excessive repetition. In particular, items 19 and 20a seem to have considerable overlap and it is difficult to understand what to report in item 23 given what has been reported in items 20, 21, and 22 without introducing excessive repetition. In the case of items 19 and 20a, we believe it is best to integrate the items (iteratively, if necessary). In the case of item 23, our suggestion for item 23a is to summarize the findings (positive and negative) that will be used to propose advice for research and practice, and to discuss any other related research.

3.6 Initial Assessment of PRISMA 2020 for Software Engineering Systematic Reviews

Our initial assessment of PRISMA 2020 as a means of addressing reporting problems in software engineering SRs is that it addresses the problems identified in Section 2.1. However, its use will need SE researchers to change both their terminology and methodology for assessing primary studies rigour and reporting systematic review limitations.

In terms of its scope, PRISMA 2020 is relevant to quantitative SRs whether or not they report meta-analysis, and to mixed-method analyses that are useful in the context of evaluating complex interventions. In addition, the guidelines for reporting meta-analysis methods in item 13 and item 20 would also apply to families of experiments. However, PRISMA 2020 will be of very limited value to SE researchers unless it is also useful for the types of SR that are more widely used in SE than quantitative SRs or meta-analyses. Thus, in Section 4, we assess the relevance of PRISMA 2020 for reporting mapping studies, and, in Section 5, we consider its relevance for reporting qualitative reviews (i.e., reviews that rely on qualitative synthesis).

4 Guidelines for Mapping Studies

This section discusses whether PRISMA 2020 is relevant for mapping studies. As shown in Table 3, Tricco et al. [17] produced the PRISMA-ScR checklist for reporting scoping reviews, based on the original PRISMA statement [16]. They define a scoping review to be a "type of knowledge synthesis that follows a systematic approach to map evidence on a topic and identify main concepts, sources and knowledge gaps". Thus, Tricco et al.'s definition implies that a *scoping review* is very similar to what we refer to as a *systematic mapping study* in software engineering. Booth et al. [75] define scoping reviews and mapping studies as different forms of review. However, looking at Booth et al's description, the main difference appears to be that a mapping study addresses a broad topic area, while a scoping review aims at assessing whether there is sufficient evidence to undertake a systematic review. Thus, findings from a mapping study might be more extensive and varied than those from a scoping review. Nonetheless, *our basic assumption* is that reporting guidelines for mapping studies address the same basic items as the guidelines for scoping reviews.

To construct the guidelines for scoping reviews, Tricco et al. started by reviewing the set of items defined in the original PRISMA. They identified five items (concerning effect sizes, synthesis, risk of bias across studies, and additional analysis) that were not applicable for scoping reviews, and two items (concerning risk of bias for primary studies) as optional. In our experience, the same restrictions apply to mapping study reports.

4.1 Methodology for Identifying Mapping Study Reporting Items

In order to check whether PRISMA 2020 covered all the issues identified in PRISMA-ScR, we assessed whether the items identified as relevant in PRISMA-ScR were included (at least at a conceptual level) among the PRISMA 2020 items. We evaluated the relationship between the items in PRISMA-ScR and PRISMA 2020 items individually, in order to ensure that we were all familiar with the standards. This was done by taking a list of all PRISMA-ScR items, identifying the item number(s) of any equivalent or related PRISMA 2020 items, and adding comments about the choice of related items.

Kitchenham integrated the initial results in a spreadsheet and then circulated the integrated assessments to the other authors. All authors revised their assessments adding any relevant comments. The revised assessments (held in spreadsheets) were returned to Kitchenham who integrated the assessments and circulated the integrated assessments for a second time. We then discussed (by e-mail) any remaining disagreements. Our aims were:

- to assess whether all items in PRISMA-ScR mapped (at least conceptually) to one or more PRISMA 2020 items,
- to identify items for which the PRISMA-ScR terminology or underlying assumptions, or the PRISMA 2020 structure, would need additional explanation before PRISMA 2020 would be usable by SE researchers reporting mapping studies,
- to identify PRISMA 2020 items that would need to have extended definitions to cover the requirements for mapping studies.

4.2 Mapping Study Reporting Item Results

The results of our assessment of mapping study reporting items is shown in Table 5 which confirms that PRISMA 2020 includes all the items needed to report a mapping study, even though the description of those items may need to be revised to cover mapping studies.

When comparing the PRISMA-ScR Checklist and PRISMA 2020, there were three terminology issues that complicated our evaluations:

1) PRISMA-ScR talks about "sources of evidence". By this we understand the authors to mean an individual primary study, since it is possible that one article or report might contain more than one primary study.
2) PRISMA-ScR talks about the "data charting process". After checking the papers cited by Tricco et al. [17] (see [76] and [77]), we found this meant the process of extracting all the variables and the textual information that were used to address the research questions from each primary study in "calibrated forms" (i.e., agreed data extraction forms). In the context of SE mapping studies we often need to classify primary studies. The

specification of the classification system(s) used would be part of the data definitions item, whereas the process of extracting the classification data would be part of the data charting process item. Data charting also differs from defining the method(s) used for data synthesis, which define how the data from each primary study will be grouped and displayed.

3) PRISMA-ScR uses the term *Synthesis of Results* to remain consistent with PRISMA [16]. However, there is a substantial difference between investigating the characteristics of scientific articles and empirical studies and synthesising the outcomes of empirical studies. In the case of synthesis, a new conclusion is produced by means such as meta-analysis, narrative synthesis, or qualitative meta-synthesis. In the case of a scoping review or mapping study, different characteristics of the context and conduct (but not the outcomes) of a set of primary studies are specified and analysed to identify subsets of primary studies with similar characteristics. We would prefer to use the term *Analysis of Study Characteristics* rather than *Synthesis of Results* for SE mapping studies.

Also, PRISMA-ScR suffers from similar problems related to iteration and repetition as PRISMA 2020 (see Section 3.5). We suggest ensuring that tables and graphs representing the answers to research questions are reported for each research question with the associated textual explanation. The Discussion section should summarize results of particular importance for researchers and include comparisons with previous research.

4.3 Mapping Study Items Discussion

In this section, we discuss some of the significant differences between a systematic review and a mapping study from the viewpoint of reporting standards.

4.3.1 The Search Process

In most respects, the search for evidence is the same for mapping studies as it is for systematic reviews. However, there is less emphasis on completeness and more emphasis on defining the search process used, and specifying any search limitations (e.g., restrictions based on language, evidence sources, or publication dates) together with a rationale for any such limitations.

4.3.2 Quality Assessment

The main simplification for mapping studies compared with quantitative SRs is that the PRISMA standard for scoping reviews (PRISMA-ScR) accepts that there will be no formal aggregation of the outcomes of primary studies. This implies that risk of bias due to synthesis is irrelevant, as is certainty assessment.

Risk of bias assessment for primary studies, which Tricco et al. [17] refer to as *critical appraisal of individual sources of evidence*, is optional, but, if undertaken, the method of assessment should be reported in the Methods section and the results of the assessment should be reported in the Results section. The fact that mapping studies do not necessarily require risk of bias assessments is consistent with Yang et al.'s observation that reviewers need to consider whether it is necessary to perform quality assessment [4].

Unlike most SE mapping studies, when conducting tertiary mapping studies that investigate SR methodology, assessment of the quality of the primary studies is often required to address mapping study research questions. Thus, in terms of PRISMA-ScR, identifying quality assessment criteria and extracting quality assessments would be regarded as a data charting process.

4.3.3 Threats to Validity

PRISMA-ScR states that the Discussion section should include a discussion of any *limitations of the scoping study review process*. The explanation of the item makes the point that because critical appraisal of individual sources of evidence is optional, the limitations section should concentrate on limitations of the scoping review process. Any deviations from scoping review guidelines or the specific scoping review protocol should be "noted along with the rationale, and a reflection on the potential effect on the results". In this context any methodological limitations defined and justified in the Methods section do not need to be discussed again, unless the researchers have observed some unanticipated problems arising from their chosen methodology.

An important implication of this is that, if there have been no deviations from the secondary review standards or the review protocol and no critical appraisal of primary studies, discussion of limitations is not necessary for mapping studies.

4.4 Mapping Study Guidelines Conclusions

As noted by Tricco et al. [17] adapting a systematic review reporting standard to provide a standard suitable for scoping reviews is primarily a process of deciding which items are irrelevant or optional and modifying some of the item definitions. In this section, we have mapped the ScR items to PRISMA 2020 and explained how the ScR items relate to SE mapping studies. Thus, we conclude that, rather than develop separate standards for mapping studies it would be preferable to extend the definitions and scope of PRISMA 2020 items to include SE mapping studies.

In our opinion, some of the terminology used in PRISMA-ScR and PRISMA 2020 is inappropriate for SE mapping studies. For SE mapping studies:

- The term *Data Charting* is misleading and the PRISMA 2020 term *Data Collection Process* is more appropriate.
- The term *Synthesis of Results* in PRISMA-ScR should be replaced by *Analysis of Study Characteristics* in any standards used for SE mapping studies.

In addition, McGowan et al. [78] point out that ongoing revisions to the PRISMA statement make it likely that authors of PRISMA-related checklists such as PRISMA-ScR will consider revising those checklists to conform with PRISMA 2020. Thus, SE researchers need to be aware that new scoping study checklists are likely to be published in the near future.

5 Reporting Qualitative Systematic Reviews

Two SE tertiary studies ([79] and [80]) have emphasised the importance of using qualitative synthesis to address SR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSE.2022.3174092, IEEE Transactions on Software Engineering

TABLE 5
Mapping PRISMA-ScR items [17] to PRISMA 2020 items

Id	Review aspect	Information Required	PRISMA item
Title			
1		Identify the report as a scoping review	1
Abstract			
2		Provide a structured summary that includes (as applicable) background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
Introduction			
3	Rationale	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	3
4	Objectives	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
Methods			
5	Protocol & Registration	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	24a, 24b
6	Eligibility criteria	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	5
7	Information Sources	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	6
8	Search	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	7
9	Selection of sources of evidence	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	8
10	Data charting process	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	9, 13b
11	Data items	List and define all variables for which data were sought and any assumptions and simplifications made.	10b
12	Critical Appraisal of individual sources of evidence	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	11
13	Synthesis of results	Describe the methods of handling and summarizing the data that were charted.	9, 13c
Results			
14	Selection of sources of evidence	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	16a, 16b
15	Characteristics of sources of evidence	For each source of evidence, present characteristics for which data were charted and provide the citations.	17
16	Critical Appraisal of sources of evidence	If done, present data on critical appraisal of included sources of evidence (see item 12).	18
17	Results of individual sources of evidence	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	19, 20a
18	Synthesis of results	Summarize and/or present the charting results as they relate to the review questions and objectives.	20b
Discussion			
19	Summary of evidence	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	23a
20	Limitations	Discuss the limitations of the scoping review process.	23c
21	Conclusions	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	23d
22	Funding	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	25, 26

systematic review research questions. Therefore, a major limitation of PRISMA 2020 as a standard for reporting SE secondary studies is that, as explicitly confirmed by Page et al. [15], it does not directly support qualitative synthesis. In this section, we introduce a group of studies that discuss guidelines for reporting qualitative synthesis (see Table 6).

Section 4 confirmed that some of the PRISMA 2020 items can be used to identify guidelines for reporting mapping studies. In addition, our initial reading of guidelines for reporting qualitative reviews showed that they conformed with the basic SR reporting concepts of defining methodologies for search and synthesis, reporting the results of using those methodologies, and discussing the finding of the synthesis. Thus, the research question addressed in this section is whether SE researchers need separate guidelines for qualitative synthesis, or whether the basic structure of PRISMA 2020 is flexible enough to support qualitative synthesis reports.

5.1 Background

In this section, we introduce issues related to qualitative research and qualitative synthesis that are necessary to understand the requirements for qualitative synthesis reporting guidelines.

5.1.1 Researcher Viewpoint in Qualitative Synthesis

For engineers, such as software engineers and computer scientists, a specific problem with using any qualitative method is the need for individual researchers to consider their personal "philosophical positioning" and its relationship to their choice of qualitative methodology. For example, Table 2 in [9] identifies five underlying philosophies, which, to our knowledge, are not frequently used in the context of software engineering research. However, the realist philosophy described in [10] takes the view that an intervention alters context, which then triggers mechanisms which produce both intended and unintended outcomes. This is consistent with a view of complex systems that is familiar to computer

TABLE 6
Guidelines for Reporting Qualitative Reviews

ID	Name	Scope	Derivation
RAMESES [10]	Realist And MEta-narrative Evidence Synthesis	Reporting the outcomes of complex interventions and adopting policy friendly approaches to evidence synthesis	Guidelines for reporting guidelines [81], excluding Delphi exercise
ENTREQ [9]	ENhancing Transparency in REporting the synthesis of Quality research	A framework for reporting the synthesis of qualitative health research	Protocol for guidelines construction [82]
GRADE-CERQual [83]	Confidence in the Evidence from Reviews of Qualitative Research	To support the use of findings from systematic reviews of qualitative evidence	Development discussed in [84] See other short papers with more detailed descriptions of different aspects of CERQUal: [84], [85], [86], [87], [88], [89]

scientists and software engineers, and seems broadly consistent with mixed-methods approaches and qualitative studies, in particular, both those investigating barriers and enablers to the adoption of complex interventions and those that aim at more sophisticated syntheses.

In trying to identify whether PRISMA 2020 items can be extended to support qualitative synthesis, we have taken a pragmatic, realist approach to specifying SE systematic reviews reporting guidelines.

5.1.2 Standards for Qualitative Synthesis

Before considering guidelines for *reporting* qualitative synthesis, it would be useful to have definitive guidelines for *performing* qualitative synthesis. The original guidelines for systematic reviews in software engineering (i.e., [5] and [6] did not mention qualitative synthesis. The more recent guidelines in [90] acknowledge the need for guidelines for qualitative reviews, but treat such reviews as being deviations from the quantitative SR guidelines, and do not provide detailed advice. Producing guidelines for performing qualitative reviews is beyond the scope of this article, but to assist readers, we identify current initiatives concerned with providing such guidelines which are summarized in Table 6.

The Cochrane Qualitative and Implementation Methods Group produced a series of papers discussing qualitative synthesis, that were published in the Journal of Epidemiology and we summarize these papers briefly in this section.

Noyes et al. [91] introduce the series of papers, each of which addresses a major topic related to qualitative synthesis. The first two papers discuss the most basic issues:

- Harris et al. [92] discuss methods for question formulation, searching, and protocol development. They point out that qualitative reviews ask "how" and "why" questions. Initial questions may be broad exploratory questions that attempt to map what is known before formulating or refining questions.
- Noyes et al. [93] discuss methods to assess methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings.

The remaining three papers in the 2018 series are particularly concerned with evaluating methods for complex interventions:

- Cargo et al. [94] discuss general issues,
- Harden et al. [67] discuss methods for integrating qualitative and quantitative implementation evidence (see Section 3.4),
- Fleming et al. [95] discuss reporting guidelines for synthesis of qualitative, implementation, and process evaluation evidence. They recommend the *ENTREQ* method [9] which is a checklist for qualitative reviews and the *RAMESES* method [10] which is a checklist for realist and meta-narrative reviews. We compare these checklists with the PRISMA items in Section 5.2.

5.1.3 GRADE-CERQual: Confidence in Syntheses of Qualitative Evidence

PRISMA 2020 requires reports of SRs to include an assessment of the confidence in any synthesis of primary study findings. The GRADE-CERQual initiative has produced a set of guidelines similar to the GRADE guidelines but aimed at qualitative synthesis, see Table 6. In this section, we summarize the content of each of the GRADE-CERQual related papers.

Lewin et al. [84] introduces the ideas and scope of the set of guidelines, then:

1) Lewis et al. [85] explain how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table.
2) Munthe-Kaas et al. [86] discuss how to assess methodological limitations of qualitative primary studies. Such limitations relate to the body of evidence that contributes to each finding reported in the qualitative synthesis. Assessments should emphasize *methodological strengths and weaknesses* rather than the quality of reporting.
3) Colvin et al. [87] discuss how to assess primary study coherence, which is defined as "an assessment of how clear and cogent is the fit between the data from the primary studies and a review finding that synthesizes that data".
4) Glenton et al. [88] discuss adequacy of data, which is somewhat analogous to sample size and number of experiments in quantitative studies. It requires an assessment of whether the number of participants and the richness of the data obtained from the participants are sufficient to understand and explain a phenomenon.
5) Noyes et al. [89] discuss the relevance of data which is defined as "the extent to which the body of data from the primary studies supporting a review finding is applicable to the context specified in the research question".

We provide a more detailed discussion of GRADE-CERQual in the Supplementary Material [8].

5.2 A Comparison of Qualitative and Quantitative Review Reporting Checklists

To investigate the differences between qualitative synthesis reporting approaches and PRISMA 2020, we first compared the items in Tong et al's ENTREQ checklist [9] to PRISMA 2020 items and subsequently compared the items in the RAMESES checklist [10] to PRISMA 2020 items.

5.2.1 Comparison Methodology

For each qualitative synthesis reporting checklist, we assessed their correspondence to PRISMA 2020 items using the same approach that we used for mapping studies (see Section 4.1).

We each assessed independently whether items in the checklist mapped (at least at a conceptual level) to the corresponding items in PRISMA 2020. Kitchenham collated the three independent assessments and circulated the collated assessments (together with any written comments we made to support our assessment). This process was repeated two times. Subsequently, we reviewed the final integrated assessment and comments and discussed our disagreements until we reached a consensus.

5.2.2 Result of comparing ENTREQ with PRISMA 2020

The comparison with the ENTREQ checklist is shown in Table 7 and confirms that all the items in the ENTREQ checklist can be mapped to PRISMA 2020 items. The main differences between the two checklists are:

- Levels of abstraction. In some cases ENTREQ items were at a lower level of detail than PRISMA 2020, so that several related ENTREQ items mapped to a single PRISMA 2020 item. For example, the methods used to assess Risk of Bias (PRISMA 2020 item 11) were covered by three items in the ENTREQ checklist (i.e., items 10, 11, 12).
- The order of the checklist items. PRISMA 2020 is designed around a general order of Introduction, Methods, Results, Discussion which relates to the order in which different activities are conducted. In contrast, ENTREQ is organized around five *domains* identified as: Introduction, Synthesis Methodology, Literature Search and Selection, Appraisal, and Synthesis of findings. The domains broadly correspond to topics that address the same concern rather than issues that are addressed within the same review process step. Thus, in the case of the final three topics, the methods used to address that topic, and the results of using the methods are discussed together.
- ENTREQ omits some standard items such as title, abstract, protocol registration, data availability, and financial support.
- Of most significance is that ENTREQ omits any discussion of publication bias or confidence in the body of evidence. Omitting any discussion of publication bias is sensible in the context of qualitative reviews since the selection process does not usually require completeness (for example, it may be based on theoretical saturation). However, it is important to ensure that there has been a thorough search of the literature to avoid missing relevant disconfirming cases [96]. In addition, there is also a need to assess the confidence in the evidence, which can be addressed using the GRADE CERQUal method, discussed in Section 5.1.3.

5.2.3 Result of comparing RAMESES with PRISMA 2020

The comparison with the RAMESES checklist is shown in Table 8 and confirms that all of the items in the RAMESES checklist mapped to items in the PRISMA 2020 checklist. There were similar issues to those identified for the ENTREQ checklist, where a single RAMESES item mapped to several different PRISMA 2020 items. In particular, selection and appraisal of documents mapped to one element, whereas in PRISMA 2020 it maps to items related to defining eligibility criteria, defining how the eligibility criteria are applied, and reporting the results of the selection process. In addition, RAMESES does not explicitly mention critical appraisal of documents as an issue that is separate from document selection. It also omits some of the elements related to scientific ethics that are included in PRISMA 2020.

RAMESES requires reviewers to comment on the strength of evidence supporting each finding. As we suggested in the context of ENTREQ, GRADE-CERQual would be a suitable method for addressing this issue.

5.2.4 Comparison Conclusion

We conclude that, although not explicitly recommended by Page et al., given appropriately extended item definitions, the ENTREQ and RAMESES checklists can be mapped to the structure and items of PRISMA 2020.

6 SEGRESS: Software Engineering Guidelines for REporting Secondary Studies

In Section 4, we assessed the relationship between PRISMA 2020 and PRISMA-ScR [17] for scoping reviews and in Section 5, we assessed the relationship between PRISMA 2020 and the ENTREQ and RAMESES checklists for reporting qualitative reviews. We concluded that PRISMA 2020 item definitions could be extended to cater for reporting scoping reviews and mapping studies and qualitative reviews. In this section, we show that the PRISMA 2020 structure is flexible enough to cater for mapping studies and qualitative reviews and demonstrate how the individual item definitions can be extended to support these forms of review.

6.1 SEGRESS Development

After assessing the requirements for SE reporting guidelines, we developed the SEGRESS checklist, shown in Table 9. SEGRESS is based on the PRISMA 2020 structure [15], but incorporates information from PRISMA-ScR [17] for mapping studies, and from ENTREQ [9] and RAMESES [10] for qualitative synthesis reviews.

For each item of the SEGRESS checklist, we identify the scope of the generic definition and provide additional comments related to the type of SR, if necessary. Requirements for mixed-methods reviews are in most cases the same as those for quantitative reviews. Readers performing tertiary studies related to assessing research methods should read the comments related to mapping studies. Researchers

TABLE 7
Mapping ENTREQ Qualitative Synthesis Items [9] to PRISMA 2020 items

Id	Review aspect	Information Required	PRISMA item
	Domain 1	**Introduction**	
1	Aim	State the research questions the synthesis addresses.	4
	Domain 2	**Method**	
2	Synthesis Methodology	Identify the synthesis methodology or the theoretical framework which underpins the synthesis and describe the rationale for choice of methodology.	13d
	Domain 3	**Literature search and selection**	
3	Approach to Searching	Indicate whether the search was pre-planned or iterative.	7
4	Inclusion criteria	Specify the inclusion criteria.	5
5	Data sources	Describe the information sources used (e.g., digital libraries), when the search was conducted and the rationale for the using the data source.	6
6	Electronic search strategy	Define search strings used.	7
7	Study screening methods	Describe the methods used to screen the studies.	8
8	Study characteristics	Present the characteristics of the included studies.	10a, 17
9	Study selection results	Identify the number of studies screened and provide reasons for study inclusion.	16
	Domain 4	**Quality Appraisal**	
10	Rationale for appraisal	Describe the rationale and approach used to appraise the selected studies or study findings.	11
11	Appraisal items	State the tools, frameworks and criteria used to appraise the studies or selected findings.	11
12	Appraisal process	Indicate whether appraisal was conducted independently by more than one reviewer and if consensus was required.	11
13	Appraisal results	Present results of quality assessment and indicate which articles, if any, were weighted/excluded and give the rationale.	18
	Domain 5	**Synthesis of findings**	
14	Data Extraction	Indicate which sections of the primary studies were analysed and how the data were extracted from the primary studies.	9
15	Software	State the computer software used, if any.	9
16	Number of reviewers	Identify who was involved in the coding and analysis.	9
17	Coding	Describe the process for coding.	13b
18	Study comparison	Describe how comparisons were made within and across studies.	13c, 13e
19	Derivation of themes	Explain whether the process of deriving themes or constructs was inductive or deductive.	13d
20	Quotations	Provide quotations from the primary studies to illustrate themes/constructs and identify whether the quotations were participant quotations or the authors interpretations.	20a, 20c
21	Synthesis output	Present rich, compelling and useful results that go beyond a summary of the primary studies.	20a, 20c, 23a, 23d

performing other types of tertiary study should consult the comments related to quantitative and qualitative reviews, as appropriate.

A limitation of SEGRESS is that the authors of ENTREQ and RAMESES both acknowledge that their checklists are preliminary checklists. This means that SE researchers must remain alert for any changes in the ENTREQ and RAMESES checklists that could require them to provide additional information when they report qualitative reviews.

6.2 Preliminary Validation of SEGRESS

The PRISMA 2020 authors did not provide any empirical validation of their checklist. However, in addition to the checklist, they provided a more detailed explanation for each item, a list of issues that need to be addressed by each item and an excerpt from a published SR related to each item. This allows readers to better appreciate the rationale for each item, what needs to be reported for each item, and how the approach can be implemented in practice.

Following their example, we provide a more detailed discussion of each item in the Supplementary Material [8], based on the explanations provided by PRISMA 2020, and present examples of how the item was reported in published software engineering SRs. We concentrate on excerpts from a variety of quantitative and qualitative SRs, at the expense of mapping studies. We made this decision because mapping studies are generally easier to report than full SRs. They do not undertake synthesis of results and have no requirement to undertake assessment of risk of bias, risk of missing values, or certainty.

Where possible, we use examples of our own SRs, but for qualitative reviews and meta-analysis, we use excerpts from the following SRs:

- For qualitative synthesis, the SR on software engineer motivation by Beecham, Sharp and their colleagues reports an SR that undertook an extensive qualitative synthesis which included a model validation exercise (see [97], [98] and [99]). For RoB and certainty in the body of evidence, we report excerpts from [100], [46] and [101].
- For meta-analysis, we have relied heavily on descriptions of the meta-analysis graphics reported by Hannay et al. [102], but for copyright reasons, we cannot use the graphics themselves. We also use excerpts from an SR that undertook statistical heterogeneity analysis [36].

A problem with the idea of providing independent excerpts for each item is that they do not provide readers with an idea of how related items interact. For this reason, we have also included a running example by revising the report of a SR undertaken by Kitchenham, Mendes and Travassos on the comparative accuracy of single company and cross-company estimation models reported in a protocol, conference paper, and journal paper ([103], [104] and [105]). This SR is an example of a quantitative SR that did not use meta-analysis. The running example acts as a trial of how the SEGRESS items for reporting study RoB, missing data RoB and certainty in the body of evidence can be integrated into an overall assessment of the quality of evidence.

TABLE 8
Mapping RAMESES Qualitative Synthesis Items [10] to PRISMA 2020 items

Id	Review aspect	Information Required	PRISMA item
Title			
1		Identify the document as a realist synthesis or review.	1
Abstract			
2		Brief details of the background, review questions or objectives, search strategy, method of selection, appraisal, analysis and synthesis, main results, and implications for practice.	2
Introduction			
3	Rationale for review	Explain why the review is needed and what it is likely to contribute to existing understanding of the topic area.	3
4	Objectives and focus of review	State the objectives of the review and/or the review question(s). Define and provide a rationale for the focus of the review.	4
Methods			
5	Changes in the review process	Any changes made to the review process that was originally planned should be briefly described and justified.	24c
6	Rationale for using realist synthesis	Explain why realist synthesis was considered the most appropriate method to use.	13d
7	Scoping the literature	Describe and justify the initial process of exploratory scoping of the literature.	7
8	Searching process	State and provide a rationale for how the iterative searching was done. Provide details of all the sources accessed for information in the review. For electronic databases report, for example, name of database, search terms, dates of coverage and date last searched. If researchers with topic knowledge were contacted, indicate how they were identified and selected.	7, 8
9	Selection and appraisal of documents	Explain how judgments were made about including and excluding data from documents, and justify these.	5, 8
10	Data Extraction	Describe and explain which data or information were extracted from the included documents and justify this selection.	10a
11	Analysis and synthesis processes	Describe the analysis and synthesis processes in detail. This section should include information on the constructs analyzed and describe the analytic process.	13
Results			
12	Document flow diagram	Provide details on the number of documents assessed for eligibility and included in the review with reasons for exclusion at each stage as well as an indication of their source of origin (for example, from searching databases, reference lists and so on).	16
13	Document characteristics	Provide information on the characteristics of the documents included in the review.	17
14	Main findings	Present the key findings with a specific focus on theory building and testing.	20a
Discussion			
15	Summary of findings	Summarize the main findings, taking into account the review's objective(s), research question(s), focus and intended audience(s).	23a
16	Strengths, limitations and future research directions	Discuss both the strengths of the review and its limitations. These should include (but need not be restricted to) (a) consideration of all the steps in the review process, and (b) comment on the overall strength of evidence supporting the explanatory insights which emerged. The limitations identified may point to areas where further work is needed.	23b, 23c
17	Comparison with existing literature	Where applicable, compare and contrast the review's findings with the existing literature (for example, other reviews) on the same topic.	23a
18	Conclusion and recommendations	List the main implications of the findings and place them in the context of other relevant literature. If appropriate, offer recommendations for policy and practice.	23d
19	Funding	Provide details of funding source (if any) for the review, the role played by the funder (if any) and any conflicts of interest of the reviewers.	25, 26

The individual item examples and the running example are reported in our Supplementary Material [8]. They provide a preliminary validation that the SEGRESS items are appropriate for qualitative reviews as well as quantitative SRs and meta-analyses. The running example demonstrates that the SEGRESS items properly address the reporting of risk of bias and quality of evidence, which are the most challenging aspects of PRISMA 2020 and SEGRESS.

6.3 Issues Arising from the Preliminary Validation of SEGRESS

As we expected, the issue of coordinating the assessments of risk of primary study bias and risk of missing data in order to produce a GRADE style assessment of the certainty in the body of evidence is the most difficult part of using the SEGRESS checklist. An advantage of the SEGRESS checklist was that thinking in terms of the GRADE and GRADE-CERQual criteria was useful for identifying potential problems with the primary studies that we did not identify in the original SR. The problem with the items related to risk of bias and certainty is that assessing these issues is subjective. The most difficult problem is assessing the risk of missing data/projects because the SR authors need to assess the rigour of their own methods. In contrast, assessing the risk of primary study bias involves the rigour of the primary study authors and assessing other GRADE criteria such as Inconsistency and Indirectness is about the characteristics of the set of primary studies results contributing to a specific finding or synthesis outcome.

Nevertheless, the examples have confirmed that the individual items identified in SEGRESS are relevant to both quantitative SRs and qualitative reviews, and that, at least some authors of software engineering SRs are currently adopting approaches consistent with SEGRESS items.

As a result of our preliminary validation, we have identified several issues that might influence decisions to adopt the SEGRESS checklist:

- We have adapted the explanation of some items of PRISMA 2020 that seem more relevant to medical SRs rather than to software engineering SRs, based on our view of best SE practice. Some researchers might dispute our view of good practice in SE and prefer to

conform strictly to the PRISMA 2020 guidelines.

- A practical concern is that conforming to SEGRESS (or indeed PRISMA 2020) may increase the length of reports of SRs. This may be acceptable if the outcome of the SR is a simple meta-analysis, but if the outcomes are more complex (such as a qualitative model that needs definition and explanation), it may cause serious length issues. Authors should consider referencing published protocols, preparing supplementary material, or publishing the SR results and any complex model building exercise in separate publications[3].
- Contrary to the order implied by PRISMA 2020, when creating our running example in the Supplementary Material [8], we thought that it made more sense to perform any sensitivity analysis which could lead to a revision of the SR analysis or synthesis *before* initiating any investigation of possible reasons for the heterogeneity of the results. We have therefore changed the order of items 13d and 13f, items 20c and 20d in SEGRESS. In general, the order in which sub-items are discussed that is adopted in a specific SR should be decided by the authors in order to support report clarity.

7 Discussion and Conclusions

In Section 2, we discussed four tertiary studies that raised concerns about the standard of SR reports in SE. We identified the need for reporting guidelines that identify all required items in the context of a well-defined report structure that limits unnecessary repetition. Two specific issues that caused particular concern were reporting threats to validity and quality assessment. However, as Budgen et al. [1] point out:

- It is important to undertake quality assessment of individual studies and use the results of that assessment constructively.
- It is important to assess the strength of evidence associated with any recommendations that are based on synthesising the available evidence.

The advantage of PRISMA 2020 is that it provides a framework for addressing these issues (see Figure 1).

The authors of PRISMA 2020 state that it is suitable for mixed-methods reviews, which we discuss in Section 3.4. Mixed-methods reviews are particularly important for industry-based interventions, when outcomes are influenced by the complex nature of the relationship between the intervention and its environment. They can be used to help researchers to interpret/explain the results of quantitative reviews, when qualitative data from the primary studies or qualitative reviews on the same topic are available.

A major disadvantage of PRISMA 2020 for software engineering use is that it is strongly oriented towards quantitative systematic reviews and meta-analysis of formal experiments and quasi-experiments, whereas secondary studies in SE are mainly qualitative studies or mapping studies [7]. Tricco et al. [17] showed that the original PRISMA statement [16] was suitable for reporting scoping studies and, in Section 4, we confirm that PRISMA 2020 is also suitable for mapping studies. However, the definitions of some items need to be extended and some items are not relevant, specifically those related to synthesising primary study outcomes and assessing the validity of synthesized findings.

The authors of PRISMA 2020 report that it is unsuitable for qualitative reviews, nevertheless, in Section 5 we show that PRISMA can be used as a framework for reporting results of qualitative synthesis by mapping the PRISMA items to two important qualitative reporting guidelines (ENTREQ [9] and RAMESES [10]). As a result of assessing PRISMA items against PRISMA-ScR, ENTREQ and RAMESES, we were able to develop the SEGRESS checklist defined in Table 9. SEGRESS relates the PRISMA 2020 items to the requirements of mapping studies and qualitative studies. For each item, we specify which types of systematic review it is relevant to and, if necessary, include specific information related to the different types of systematic review. We hope that SE researchers interested in mixed-methods and qualitative synthesis will trial the SEGRESS checklist and comment on their experiences.

We have provided an preliminary validation of SEGRESS in the Supplementary Material [8], based on examples from the SE literature and our running example. We hope this is sufficient to encourage the SE community to undertake more extensive empirical validation through use, particularly for qualitative reviews. Adoption of SEGRESS presents a greater risk for qualitative reviews than quantitative reviews and mapping studies. SEGRESS is based on PRISMA 2020, so it is well-suited to quantitative reviews and mapping studies (which use a subset of the quantitative review items). To support the use of SEGRESS for qualitative reviews, we provide an introduction to current standards for qualitative synthesis in Section 5, together with SE examples and a detailed discussion of GRADE-CERQual in the Supplementary Material [8]. However, SE researchers need to be aware that new versions of PRISMA-ScR and the standards for reporting qualitative reviews are likely to be published in the near future. However, any risks associated with adopting SEGRESS need to be balanced against the SR reporting problems identified in software engineering. Currently, many aspects of reporting practice are being criticized, but there is no holistic view of the requirements for reporting SRs that can help researchers decide what should be done without a risk of introducing other problems. SEGRESS attempts to address this issue and provide an overall integrated framework to support the reporting of all types of SE systematic reviews.

3. It is generally accepted that full reports of systematic reviews result in long documents, and other disciplines have identified a variety of methods to address this, such as the 1-3-25 method [106]. This approach advocates a one-page summary of the findings formulated as a set of "take-home" messages, aimed at end-users of the evidence rather than researchers, a three-page executive summary of the study and its findings, addressing the needs of the sponsor and policy-makers, and a 25-page detailed report on the design and conduct of the study that is intended for reviewers and others who need to know how the review was conducted. However, such initiatives have not yet been adopted in SE.

TABLE 9. SEGRESS: The PRISMA 2020-inspired structured checklist for reporting SE secondary studies

Section	PRISMA Item	Description
Full Report		*Use of SEGRESS may result in long documents. For publication purposes, authors should consider referencing material in the protocol, publishing some material in supplementary material, and reporting any large-scale model building exercise separately from the basic SR report.*
Title		**Identify both the report topic and type of secondary study, so potential readers can find the report.**
Title	1	Identify the report as a systematic review, systematic mapping study, tertiary study, qualitative review, or mixed-methods review and specify the topic being reviewed, see explanation and examples in [8, Sec. 2.1]. Required for all review types.
Abstract		**Provide a summary of the entire report, so potential readers can easily assess its relevance.**
Structured abstract	2	Provide a structured summary incl.: Background (emphasizing the importance of this research), Objective, Methods, Results, Limitations (optional), Conclusion. Guidelines for constructing an abstract can be found in [15, Table 2] and [20, Box 2] and are discussed in the SEGRESS Supplementary Material [8, Sec. 2.2]. Required for all review types.
Introduction		**Set context for the work.**
Opening		Introduce the larger problem the paper is targeting, lay out a broad context for the work, and highlight the importance of the work to a large audience. In subsequent steps define the research area, establish a niche within the area (knowledge gap), and then focus on the niche.
Rationale	3	Describe information the reader needs to understand the work the authors did, why it is important, i.e., the rationale for the study (e.g., update, new topic area, new empirical results, mature topic having no previous systematic review) and how it contributes to the larger problem, see explanation and example in [8, Sec. 2.3]. Required for all review types.
Objectives	4	Specify the research questions, explaining how they contribute to the larger problem, see [8, Sec. 2.4]. Required for all review types.
Methods		**Outline procedures you followed and resources you used to conduct your work.**
Eligibility criteria	5	Use the study characteristics to define eligibility criteria based on the intervention or topic of interest [8, Sec. 3.1]. Criteria used to restrict the search must be specified and justified (e.g., search start and end dates, language limitations, journal restrictions, publication restrictions). Specify how any existing systematic reviews and/or qualitative reviews on the topic of interest, found by the search process, will be used. Required for all review types. Tertiary mapping studies investigating research trends must justify search restrictions, such as limiting inclusion to papers in high quality journals, in terms of the study RQs.
Information sources	6	Describe all information sources, databases, primary study references, and others (e.g., researchers) with search end dates. The Supplementary Material [8, Sec. 3.4] includes a checklist for reporting the search process based on the PRISMA-S guide [21], while [107] guides on how should software engineering secondary studies include grey material. Required for all review types.
Search Strategy	7	Present full search strategy, including, as appropriate, electronic search strings, snowballing, manual search, finding unpublished materials, and any method(s) used to assess achieved completeness. If previous reviews exist, explain how they have contributed to the current search process. The Supplementary Material [8, Sec. 3.5] includes a checklist for reporting the search process based on the PRISMA-S guide [21]. Required for all review types. Qualitative reviews should explain any search processes aimed at finding deviant cases and exceptions and any exploratory scoping of the literature.
Selection Process	8	State the process for selecting studies, including the specific phases of the selection process, the number of assessors per study, methods of handling disagreements, any tools used, and any methods of assessing agreement rates [8, Sec. 3.6]. Required for all review types. Qualitative studies should explain exclusions that relate to synthesis issues rather than eligibility criteria.
Data Collection Process	9	Specify the method used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and, if applicable, details of automation tools used in the process [8, Sec. 3.7]. Required for all review types. For qualitative reviews, indicate which areas of each primary study were analysed.
Data items	10a	List, define and justify all outcomes for which data was sought, explaining their relationship to the research questions [8, Sec. 3.8]. Required for all review types except Mapping studies, because they do not analyse primary study outcomes.
	10b	List and define all non-outcome variables for which data was sought (e.g., participant and intervention characteristics, funding source). Describe any assumptions made about any missing or unclear information [8, Sec. 3.9]. Required for all review types. For mapping studies define any classification systems used to categorize the data items and confirm how the data item relates to the research questions.
Study Risk Of Bias Assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and, if applicable, details of automation tools used in the process [8, Sec. 3.10]. This is optional for mapping studies, but required for all other review types.
Effect Measures	12	Specify for each outcome the effect measure(s) (e.g., risk ratio, mean difference) used in the synthesis or presentation of results [8, Sec. 3.11]. This is required for quantitative reviews and meta-analyses. It is sometimes reported by mapping studies, depending on the research questions (e.g., if the research question involves identifying the definitions of outcome metrics used in empirical studies). It is not required for qualitative reviews.
Analysis and Synthesis methods	13	Quantitative SRs and qualitative reviews should report the methods used for synthesis of primary study outcomes [8, Sec. 3.12]. Mapping studies should report the methods used to analyse primary study characteristics.
	13a	Describe the process used to decide which studies were eligible for each synthesis [8, Sec. 3.13].
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling missing summary statistics, or data conversions [8, Sec. 3.14]. Not required for mapping studies. Qualitative studies should describe the coding processes adopted and specify whether it was inductive (i.e., based on deriving the code from the raw textual data, which is typical for grounded theory analyses), or deductive (i.e., based on pre-existing themes or theories).
	13c	Describe any methods used to tabulate or visually display results of individual studies and synthesis [8, Sec. 3.15]. Required for all review types. For mapping studies describe the methods used to prepare tables, graphs and maps of study characteristics.
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s) [8, Sec. 3.16]. Required for all types of review except mapping studies. If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of heterogeneity, and the software packages(s) used. Qualitative studies should, where necessary, identify constructs analyzed, explain how findings from different studies were compared, and specify how synthesized findings were validated.

Continued on next page

Section	PRISMA Item	Description
	13e	Describe any sensitivity analysis conducted to assess robustness of the synthesized results [8, Sec. 3.17]. Formal procedures are available for quantitative synthesis and mixed-methods analysis, such as removing high influence data points. For qualitative methods, this involves discussing the impact of any deviant cases and exceptions on the synthesized findings. Not required for mapping studies.
	13f	Describe any methods used to explore possible causes of heterogeneity among study results [8, Sec. 3.18]. Required for all types of review except mapping studies.
Reporting Bias Assessment	14	Describe any methods used to assess risk of bias due to publication bias [8, Sec. 3.19]. Not required for mapping studies, or secondary studies investigating SE research practices rather than SE development and maintenance methods.
Certainty Assessment	15	Describe methods used to assess certainty (or confidence) in the body of evidence for an outcome (e.g., GRADE) [8, Sec. 3.20]. Not required for mapping studies or secondary studies investigating SE research practices, but essential for all other review types. See Section 3.3.3 and Section 5.1.3.
Results		**Communicate complex, quantitative and qualitative information in an easy to read manner.**
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram [8, Sec. 4.1]. Report agreement statistics, if collected. Required for all review types. Qualitative studies should describe any iteration between selection and synthesis.
	16b	Cite studies that met many but not all inclusion criteria ('near-misses') and explain why they were excluded [8, Sec. 4.2]. Optional for mapping studies, required for all other review types. Qualitative reviews should identify any eligible studies that were excluded from synthesis and justify the exclusions.
	[17-22]	*Reporting Style: If reporting syntheses (i.e., meta-analysis results or answers to research questions) obtained from different subgroups of primary studies or different research questions consider using an iterative reporting approach, keeping items 17 to 22 together for primary studies subgroups or specific research questions. Note that, even if using an iterative style for reporting, it may be appropriate to report information that was obtained from every primary study in integrated tables. The issue is that risk of bias among contributing primary studies will be different for different syntheses if they depend on different subsets of studies.*
Study characteristics	17	Describe the characteristics of each included study, and provide citations [8, Sec. 4.3]. Required for all review types.
Risk of Bias in Studies	18	Present data on the risk assessment for each study [8, Sec. 4.4]. Report agreement statistics. Optional for mapping studies but required for all other review types.
Results of individual studies	19	For quantitative reviews, for all outcomes, present for each study [8, Sec. 4.5]: a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g., confidence/credible interval), ideally using structure tables or plots. For qualitative reviews, present the major findings from each study included in the synthesis. Not usually required for mapping studies.
Results of Analyses and Syntheses	20	Quantitative SRs and Qualitative reviews should describe the results of their syntheses [8, Sec. 4.6]. Mapping studies should report their analyses of primary study characteristics.
	20a	Report each synthesis, briefly summarising the characteristics and risk of bias among contributing studies [8, Sec. 4.7]. Required for all review types. For qualitative studies, define any derived themes, and focus on theory building and testing. Provide appropriate quotations specifying the primary study from which the quotation was obtained, and whether it was produced by the study authors or individual study participants. For mapping studies, discuss the maps and tables produced to address each research question.
	20b	Present results of all statistical syntheses conducted [8, Sec. 4.8]. If meta-analysis was performed, present for each analysis, the summary estimate and its precision (e.g., confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect. Only required for quantitative reviews.
	20c	Present results of all sensitivity analysis conducted to assess the robustness of the synthesized results [8, Sec. 4.9]. Qualitative studies should discuss deviant cases and exceptions [96] and should report any additional validation of qualitative models.
	20d	Present results of all investigations of possible causes of heterogeneity among study results [8, Sec. 4.10]. Not required for mapping studies. Other review types should attempt to identify qualitative factors that might explain different primary study outcomes.
Reporting Biases	21	Report results of assessing publication bias for each synthesis [8, Sec. 4.11]. For meta-analysis, report the heterogeneity among studies and provide funnel plots. Not usually required for mapping studies or qualitative studies.
Certainty of Evidence	22	Present assessment of certainty (or confidence) in the body of evidence for each reported finding [8, Sec. 4.12]. Not required for mapping studies. Required for all other review types.
Discussion		**Turn data into knowledge (i.e., advice or recommendations for practitioners, academics, and educators), point out how your results provide novel understanding, challenge previous knowledge, or resolve persisting controversy answering questions raised in the Introduction.**
Discussion	23a	Provide a general interpretation of the results in the context of other evidence [8, Sec. 5.2]. Where applicable compare review findings with other reviews on the same topic. Required for all review types.
	23b	Discuss any limitations of the evidence included in the review [8, Sec. 5.3]. Required for quantitative and qualitative reviews. Not required for mapping studies.
	23c	Discuss any limitations of the review process used [8, Sec. 5.4]. Required for all reviews, but include only those issues that were not previously addressed as part of the specification of the specified review process or when discussing the synthesis results.
	23d	Discuss implications of the results for practice, policy and future research [8, Sec. 5.5]. Required for all review types. For mapping studies, only discussion of future research is relevant.
Registration and Protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered [8, Sec. 6.2]. Guidelines for constructing an SR protocol can be found in the PRISMA-P statement [18]. Optional for all review types.
	24b	Indicate where the review protocol can be accessed or state why no protocol is available [8, Sec. 6.3]. Optional for mapping studies, required for all other review types.
	24c	Describe and explain any amendments to information provided at registration or in the protocol [8, Sec. 6.4]. Required for quantitative and qualitative review types, optional for mapping studies.
Support	25	Describe sources of financial and non-financial support for the review and the role of the funders or sponsors of the review [8, Sec. 6.5]. Required for all review types.
Competing Interests	26	Declare competing interests of the review authors [8, Sec. 6.6]. Required for all review types.
Availability Of Data, Code and Other Materials	27	Report which of the following are publicly available and where they can be found (e.g., Zenodo, Figshare, Dryad): template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review or to produce the review (e.g., Rnw file if using R scripts or code chunks as analytic code) [8, Sec. 6.7] [108]. Optional but recommended for all review types.

REFERENCES

[1] D. Budgen, P. Brereton, S. Drummond, and N. Williams, "Reporting systematic reviews: Some lessons from a tertiary study," *Information and Software Technology*, vol. 95, pp. 62—74, 2018.

[2] X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang, "A map of threats to validity of systematic literature reviews in software engineering," in *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2016, pp. 153–160.

[3] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," *Information and Software Technology*, vol. 106, pp. 201 – 230, 2019.

[4] L. Yang, H. Zhang, H. Shen, X. Huang, X. Zhou, G. Rong, and D. Shao, "Quality assessment in systematic literature reviews: A software engineering perspective," *Information and Software Technology*, vol. 130, no. 106397, 2021.

[5] B. Kitchenham, "Procedures for undertaking systematic reviews," Keele University, UK, Tech. Rep., 2004.

[6] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Joint Technical Report Keele and Durham Universities, UK, Tech. Rep., 2007.

[7] D. Budgen and P. Brereton, "Short communication: Evolution of secondary studies in software engineering," *Information & Software Technology*, p. 106840, 2022.

[8] B. Kitchenham, L. Madeyski, and D. Budgen, "Supplementary Material for SEGRESS: Software Engineering Guidelines for REporting Secondary Studies," 2022. [Online]. Available: https://madeyski.e-informatyka.pl/download/SEGRESS22supplement.pdf

[9] A. Tong, K. Flemming, E. McInnes, S. Oliver, and J. Craig, "Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ," *BMC Medical Research Methodology*, vol. 12, no. 181, 2012.

[10] G. Wong, T. Greenhalgh, G. Westhorp, J. Buckingham, and R. Pawson, "RAMESES publication standards: realist syntheses," *BMC Medicine*, vol. 11, no. 21, 2013.

[11] M. J. Page, D. Moher, P. M. Bossuy, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. M. Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. T. andAndrea C Tricco, V. A. Welch, P. Whiting, and J. E. McKenzie, "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *MetaArXiv*, 2020.

[12] D. Budgen, P. Brereton, N. Williams, and S. Drummond, "What support do systematic reviews provide for evidence-informed teaching about software engineering practice?" *e-Informatica Software Engineering Journal*, vol. 14, pp. 7—60, 2020. [Online]. Available: https://doi.org/10.37190/e-Inf200101

[13] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Design for Generalized Causal Inference*. Houghton Mifflin Company, 2002.

[14] Y. Zhou, H. Zhang, X. Huang, S. Yang, M. Babar, and H. Tang, "Quality assessment of systematic reviews in software engineering: A tertiary study." in *Proceedings of EASE '15*, ser. 19th International Conference of Evaluation an Assessment in Software Engineering. ACM, 2015.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n71

[16] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ*, vol. 339, 2009.

[17] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, and S. E. Straus, "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Annals of internal medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[18] L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. L. (deceased), M. Petticrew, P. Shekelle, and L. A. Stewart, "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation," *BMJ Research methods and Reporting*, 2015.

[19] E. M. Beller, P. P. Glasziou, D. G. Altman, S. Hopewell, H. Bastian, I. Chalmers, P. C. Gøtzsche, T. Lasserson, and D. Tovey, "PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts," *PLOS Medicine*, vol. 10, no. 4, 2013.

[20] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and J. E. McKenzie, "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n160

[21] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, H. Blunt, T. Brigham, S. Chang, J. Clark, A. Conway, R. Couban, S. de Kock, K. Farrah, P. Fehrmann, M. Foster, S. A. Fowler, J. Glanville, E. Harris, L. Hoffecker, J. Isojarvi, D. Kaunelis, H. Ket, P. Levay, J. Lyon, J. McGowan, M. H. Murad, J. Nicholson, V. Pannabecker, R. Paynter, R. Pinotti, A. Ross-White, M. Sampson, T. Shields, A. Stevens, A. Sutton, E. Weinfurter, K. Wright, S. Young, and P.-S. Group, "PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews," *Systematic Reviews*, vol. 10, no. 1, p. 39, 2021.

[22] A. D. Oxman, "Grading quality of evidence and strength of recommendations," *BMJ*, vol. 328, pp. 1490–1494, 2004.

[23] G. H. Guyatt, A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H. J. Schünemann, "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations," *British Medical Journal*, vol. 336, pp. 924–926, 2008.

[24] G. Guyatt, A. D. Oxman, E. A. Akl, R. Kunz, G. Vist, J. Brozek, S. Norris, Y. Falck-Ytter, P. Glasziou, H. deBeer, R. Jaeschke, D. Rind, J. Meerpohl, P. Dahm, and H. J. Schünemann, "Grade guidelines: 1. introduction—grade evidence profiles and summary of findings tables," *Journal of clinical epidemiology*, vol. 64, no. 4, pp. 383–394, 2011.

[25] G. Guyatt, A. D. Oxman, R. Kunz, D. Atkins, J. Brozek, G. Vist, P. Alderson, P. Glasziou, Y. Falck-Ytter, and H. J. Schünemann, "GRADE guidelines: 2. Framing the question and deciding on important outcomes," *Journal of clinical epidemiology*, vol. 64, pp. 395–400, 2011.

[26] H. Balshem, M. Helfand, H. J. Schünemann, A. D. Oxman, R. Kunz, J. Brozek, G. E. Vist, Y. Falck-Ytter, J. Meerpohl, S. Norris, and G. H. Guyatt, "GRADE guidelines: 3. Rating the quality of evidence," *Journal of clinical epidemiology*, vol. 64, no. 4, pp. 401–406, 2011.

[27] G. H. Guyatt, A. D. Oxman, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, V. Montori, E. A. Akl, B. Djulbegovic, Y. Falck-Ytter, S. L. Norris, J. W. Williams, D. Atkins, J. Meerpohl, and H. J. Schünemann, "GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias)," *Journal of clinical epidemiology*, vol. 64, no. 4, pp. 407–415, 2011.

[28] G. H. Guyatt, A. D. Oxman, V. Montori, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, B. Djulbegovic, D. Atkins, Y. Falck-Ytter, J. W. Williams, J. Meerpohl, S. L. Norris, E. A. Akl, and H. J. Schünemann, "GRADE guidelines: 5. Rating the quality of evidence—publication bias," *Journal of clinical epidemiology*, vol. 64, no. 12, pp. 1277–1282, 2011.

[29] G. H. Guyatt, A. D. Oxman, R. Kunz, J. Brozek, P. Alonso-Coello, D. Rind, P. Devereaux, V. M. Montori, B. Freyschuss, G. Vist, R. Jaeschke, J. W. Williams, M. H. Murad, D. Sinclair, Y. Falck-Ytter, J. Meerpohl, C. Whittington, K. Thorlund, J. Andrews, and H. J. Schünemann, "GRADE guidelines: 6. Rating the quality of evidence—imprecision," *Journal of clinical epidemiology*, vol. 64, no. 12, pp. 1283–1293, 2011.

[30] G. H. Guyatt, A. D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, P. Glasziou, R. Jaeschke, E. A. Akl, S. Norris, G. Vist, P. Dahm, V. K. Shukla, J. Higgins, Y. Falck-Ytter, and H. J. Schünemann, "GRADE guidelines: 7. Rating the quality

of evidence—inconsistency," *Journal of clinical epidemiology*, vol. 64, no. 12, pp. 1294–1302, 2011.

[31] G. H. Guyatt, A. D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, Y. Falck-Ytter, R. Jaeschke, G. Vist, E. A. Akl, P. N. Post, S. Norris, J. Meerpohl, V. K. Shukla, M. Nasser, and H. J. Schünemann, "GRADE guidelines: 8. Rating the quality of evidence—indirectness," *Journal of clinical epidemiology*, vol. 64, no. 12, pp. 1303–1310, 2011.

[32] G. H. Guyatt, A. D. Oxman, S. Sultan, P. Glasziou, E. A. Akl, P. Alonso-Coello, D. Atkins, R. Kunz, J. Brozek, V. Montori, R. Jaeschke, D. Rind, P. Dahm, J. Meerpohl, G. Vist, E. Berliner, S. Norris, Y. Falck-Ytter, M. H. Murad, and H. J. Schünemann, "GRADE guidelines: 9. Rating up the quality of evidence," *Journal of clinical epidemiology*, vol. 64, no. 12, pp. 1311–1316, 2011.

[33] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. Hoffmann, C. D. Mulrow, L. Shamseer, and D. Moher, "Mapping of reporting guidance for systematic reviews and meta-analyses generated a comprehensive item bank for future reporting guidelines," *Journal of clinical epidemiology*, vol. 118, pp. 60–68, 2020.

[34] M. J. Page, J. E. McKenzie, P. M. Bossuy, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, and D. Moher, "Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement," MetaArXiv Preprints, Tech. Rep., 2020.

[35] J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, and V. Welch, Eds., *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley-Blackwell, 2019, version 6.0. [Online]. Available: https://www.training.cochrane.org/handbook

[36] M. Ciolkowski, "What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering," in *ESEM'09 Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 133–144. [Online]. Available: http://dx.doi.org/10.1109/ESEM.2009.5316026

[37] M. Shepperd, D. Bowes, and T. Hall, "Researcher bias: The use of machine learning in software defect prediction," *IEEE Transactions on Software Engineering*, vol. 40, no. 6, pp. 603–616, 2014.

[38] B. Kitchenham, A. Kitchenham, and J. Fellows, "The effects of inspections on software quality and productivity," *ICL Technical Journal*, pp. 112–122, May 1986.

[39] A. Sobel and M. Clarkson, "Formal methods application: an empirical tale of software development," *IEEE Transactions on Software Engineering*, vol. 28, no. 3, pp. 308–320, 2002.

[40] F. Lanubile and G. Visaggio, "Evaluating defect detection techniques for software requirements inspections," ISERN-00-08, Tech. Rep., 2000.

[41] L. Madeyski, *Test-Driven Development: An Empirical Evaluation of Agile Practice*. (Heidelberg, London, New York): Springer, 2010. [Online]. Available: https://doi.org/10.1007/978-3-642-04288-1

[42] L. Madeyski and Ł. Szała, "The impact of test-driven development on software development productivity — an empirical study," in *Software Process Improvement*, ser. Lecture Notes in Computer Science, P. Abrahamsson, N. Baddoo, T. Margaria, and R. Messnarz, Eds. Springer Berlin Heidelberg, 2007, vol. 4764, pp. 200–211. [Online]. Available: https://doi.org/10.1007/978-3-540-75381-0_18

[43] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, 2002.

[44] M. Jørgensen, T. Dybå, K. Liestøl, and D. I.K.Sjøberg, "Incorrect results in software engineering experiments: How to improve research practices," *The Journal of Systems and Software*, vol. 116, pp. 133–145, 2016.

[45] B. Kitchenham, L. Madeyski, and P. Brereton, "Problems with Statistical Practice in Human-Centric Software Engineering Experiments," in *Proceedings of the Evaluation and Assessment on Software Engineering*, ser. EASE '19. New York, NY, USA: ACM, 2019, pp. 134–143. [Online]. Available: https://doi.org/10.1145/3319008.3319009

[46] T. Dybå and T. Dingsøyr, "Strength of evidence in systematic reviews in software engineering," in *Empirical Software Engineering and Metrics (ESEM)*, 2008, pp. 179–187.

[47] E. Arisholm, H. Gallis, T. Dyba, and D. I. Sjøberg, "Evaluating pair programming with respect to system complexity and programmer expertise," *IEEE Transactions on Software Engineering*, vol. 33, no. 2, pp. 65–86, 2007.

[48] E. Arisholm and D. I. Sjøberg, "Evaluating the effect of a delegated versus centralized control style on the maintainability of object-oriented software," *IEEE Transactions on Software Engineering*, vol. 30, no. 8, pp. 521–534, 2004.

[49] M. Höst, C. Wohlin, and T. Thelin, "Experimental Context Classification: Incentives and Experience of Subjects," in *ICSE'05: International Conference on Software Engineering*. New York, NY, USA: ACM Press, 2005, pp. 470–478.

[50] M. Felderer and G. H. Travassos, *Contemporary Empirical Methods in Software Engineering*. Springer, 2020.

[51] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," in *Guide to Advanced Empirical Software Engineering*. Springer, 2007, ch. 10.

[52] K.-J. Stol and B. Fitzgerald, "The abc of software engineering research," *ACM Transactions on Software Engineering and Methodology*, vol. 27, no. 3, pp. 11:1–11:51, Sep. 2018. [Online]. Available: http://doi.acm.org/10.1145/3241743

[53] T. Menzies and M. Shepperd, ""bad smells" in software analytics papers," *Information and software technology*, vol. 112, pp. 35–47, 2019.

[54] M. Shepperd, Q. Song, Z. Sun, and C. Mair, "Data Quality: Some Comments on the NASA Software Defect Datasets," *IEEE Transactions on Software Engineering*, vol. 39, no. 9, pp. 1208–1215, 2013.

[55] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson, "Reflections on the NASA MDP data sets," *IET software*, vol. 6, no. 6, pp. 549–, 2012.

[56] T. Foss, E. Stensrud, I. Myrtveit, and B. Kitchenham, "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Transactions on Software Engineering*, vol. 29, no. 11, pp. 985–995, 2003.

[57] I. Myrtveit, E. Stensrud, and M. Shepperd, "Reliability and validity in comparative studies of software prediction models," *IEEE Transactions on Software Engineering*, vol. 31, no. 5, pp. 380–391, 2005.

[58] I. Myrtveit and E. Stensrud, "Validity and reliability of evaluation procedures in comparative studies of effort prediction models," *Empirical Software Engineering*, vol. 17, pp. 23–33, 2012.

[59] M. J. Page, J. E. McKenzie, and J. P. T. Higgins, "Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review." *BMJ Open*, vol. 8, no. 019703, 2018.

[60] M. Turner, B. Kitchenham, P. Brereton, S. Charters, and D. Budgen, "Does the technology acceptance model predict actual use? a systematic literature review," *Information and Software Technology*, vol. 52, pp. 463–479, 2010.

[61] J. Noyes, A. Booth, M. Cargo, K. Flemming, A. Harden, J. Harris, R. Garside, K. Hannes, T. Pantoja, and J. Thomas, *Cochrane Handbook for Systematic Reviews of Interventions 6.1 Chapter 21: Qualitative evidence*, J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, and V. Welch, Eds. Cochrane, 2020. [Online]. Available: https://www.training.cochrane.org/handbook

[62] E. F. France, M. Cunningham, N. Ring, I. Uny, E. A. Duncan, R. G. Jepson, M. Maxwell, R. J. Roberts, R. L. Turley, A. Booth, N. Britten, K. Flemming, I. Gallagher, R. Garside, K. Hannes, S. Lewin, G. W. Noblit, C. Pope, J. Thomas, M. Vanstone, G. M. A. Higginbottom, and J. Noyes, "Improving reporting of meta-ethnography: The eMERGe reporting guidance," *Psycho-oncology (Chichester, England)*, vol. 28, no. 3, pp. 447–458, 2019.

[63] B. Curtis, H. Krasner, and N. Iscoe, "A Field Study of the Software Design Process for Large Systems," *Communications of the ACM*, vol. 31, no. 11, pp. 1268–1287, 1988.

[64] T. K. Abdel-Hamid and S. E. Madnick, "Lessons learned from modeling the dynamics of software development," *Commun. ACM*, vol. 32, no. 12, pp. 1426–1438, Dec. 1989. [Online]. Available: https://doi.org/10.1145/76380.76383

[65] M. M. Lehman and R. Juan F, "Rules and tools for software evolution planning and management," *Annals of Software Engineering*, vol. 11, no. 1, pp. 15—44, 2001.

[66] T. Gorschek and K. Wnuk, *Contemporary Empirical Methods in Software Engineering*. Cham: Springer International Publishing, 2020, ch. Third Generation Industrial Co-production in Software Engineering, pp. 503–525. [Online]. Available: https://doi.org/10.1007/978-3-030-32489-6_18

[67] A. Harden, J. Thomas, M. Cargo, J. Harris, T. Pantoja, K. Flemming, A. Booth, R. Garside, K. Hannes, and J. Noyes, "Cochrane Qualitative and Implementation Methods Group guidance series

paper 5: methods for integrating qualitative and implementation evidence within intervention effectiveness reviews," *Journal of Clinical Epidemiology*, vol. 97, pp. 70—78, 2018.

[68] M. Jørgensen, "Forecasting of software development work effort: Evidence on expert judgement and formal models," *International Journal of Forecasting*, vol. 23, no. 3, pp. 449–462, 2007.

[69] B. Kitchenham, E. Mendes, and G. H. Travassos, "Cross versus Within-Company Cost Estimation Studies: A Systematic Review," *IEEE Transactions on Software Engineering*, vol. 33, no. 5, pp. 316—329, 2007.

[70] P. S. M. dos Santos and G. H. Travassos, "Structured synthesis method: The evidence factory tool," in *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2017, pp. 480–481.

[71] S. Martinez-Fernandez, P. S. Medeiros Dos Santos, C. P. Ayala, X. Franch, and G. H. Travassos, "Aggregating empirical evidence about the benefits and drawbacks of software reference architectures," in *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2015, pp. 1–10.

[72] W. A. Chapetta and G. H. Travassos, "Towards an evidence-based theoretical framework on factors influencing the software development productivity," *Empirical Software Engineering*, vol. 25, no. 5, pp. 3501–3543, 2020. [Online]. Available: https://doi.org/10.1007/s10664-020-09844-5

[73] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, no. 12, pp. 2049 – 2075, 2013.

[74] D. Budgen, A. J. Burn, O. P. Brereton, B. A. Kitchenham, and R. Pretorius, "Empirical evidence about the UML: a systematic literature review," *Software: Practice and Experience*, vol. 41, no. 4, pp. 363—392, 2011.

[75] A. Booth, D. Papaioannou, and A. Sutton, *Systematic Approaches to a Successful Literature Review*. Sage, 2012.

[76] H. Arksey and L. O'Malley, "Scoping studies: towards a methodological framework," *International Journal of Social Research Methodology: Theory and Practice*, vol. 8, no. 1, pp. 19–32, 2005.

[77] D. Levac, H. Colquhoun, and K. K. O'Brien, "Scoping studies: advancing the methodology," *Implementation Science*, vol. 5, 2010.

[78] J. McGowan, S. Straus, D. Moher, E. V. Langlois, K. K. O'Brien, T. Horsley, A. Aldcroft, W. Zarin, C. M. Garitty, S. Hempel, E. Lillie, O. Tunçalp, and A. C. Tricco, "Reporting scoping reviews—PRISMA ScR extension," *Journal of Clinical Epidemiology*, vol. 123, pp. 177–179, 2020. [Online]. Available: https://doi.org/10.1016/j.jclinepi.2020.03.016

[79] D. S. Cruzes and T. Dybå, "Research synthesis in software engineering: A tertiary study," *Information and Software Technology*, vol. 53, no. 5, pp. 440 – 455, 2011.

[80] X. Huang, H. Zhang, X. Zhou, M. A. Babar, , and S. Yang, "Synthesizing qualitative research in software engineering: A critical review," in *Proceedings of ACM/IEEE 40th International Conference on Software Engineering*. International Conference on Software Engineering, 2018.

[81] D. Moher, K. F. Schulz, I. Simera, and D. G. Altman, "Guidance for developers of health research reporting guidelines," *PLOS Medicine*, vol. 7, no. 2, pp. 1–9, 02 2010. [Online]. Available: https://doi.org/10.1371/journal.pmed.1000217

[82] T. Greenhalgh, G. Wong, G. Westhorp, and R. Pawson, "Protocol - realist and meta-narrative evidence synthesis: Evolving standards (rameses)," *BMC Medical Research Methodology*, vol. 11, no. 1, p. 115, 2011. [Online]. Available: https://doi.org/10.1186/1471-2288-11-115

[83] S. Lewin, C. Glenton, H. Munthe-Kaas, B. Carlsen, C. J. Colvin, M. Gülmezoglu, J. Noyes, A. Booth, R. Garside, and A. Rashidian, "Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual)," *PLoS medicine*, vol. 12, no. 10, p. e1001895, 2015.

[84] S. Lewin, A. Booth, C. Glenton, H. Munthe-Kaas, A. Rashidian, M. Wainwright, M. A. Bohren, Ö. Tunçalp, C. J. Colvin, R. Garside, B. Carlsen, E. V. Langlois, and J. Noyes, "Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series," *Implementation science : IS*, vol. 13, no. Suppl 1, pp. 2–2, 2018.

[85] S. Lewin, M. Bohren, A. Rashidian, H. Munthe-Kaas, C. Glenton, C. J. Colvin, R. Garside, J. Noyes, A. Booth, Ö. Tunçalp, M. Wainwright, S. Flottorp, J. D. Tucker, and B. Carlsen, "Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 2: how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table," *Implementation science : IS*, vol. 13, no. Suppl 1, pp. 10–10, 2018.

[86] H. Munthe-Kaas, M. A. Bohren, C. Glenton, S. Lewin, J. Noyes, Ö. Tunçalp, A. Booth, R. Garside, C. J. Colvin, M. Wainwright, A. Rashidian, S. Flottorp, and B. Carlsen, "Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 3: how to assess methodological limitations," *Implementation science : IS*, vol. 13, no. Suppl 1, pp. 9–9, 2018.

[87] C. J. Colvin, R. Garside, M. Wainwright, H. Munthe-Kaas, C. Glenton, M. A. Bohren, B. Carlsen, Ö. Tunçalp, J. Noyes, A. Booth, A. Rashidian, S. Flottorp, and S. Lewin, "Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 4: how to assess coherence," *Implementation science : IS*, vol. 13, no. Suppl 1, pp. 13–13, 2018.

[88] C. Glenton, B. Carlsen, S. Lewin, H. Munthe-Kaas, C. J. Colvin, Ö. Tunçalp, M. A. Bohren, J. Noyes, A. Booth, R. Garside, A. Rashidian, S. Flottorp, and M. Wainwright, "Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 5: how to assess adequacy of data," *Implementation science : IS*, vol. 13, no. Suppl 1, pp. 14–14, 2018.

[89] J. Noyes, A. Booth, S. Lewin, B. Carlsen, C. Glenton, C. J. Colvin, R. Garside, M. A. Bohren, A. Rashidian, M. Wainwright, Ö. Tuncalp, J. Chandler, S. Flottorp, T. Pantoja, J. D. Tucker, and H. Munthe-Kaas, "Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 6: how to assess relevance of the data," *Implementation science : IS*, vol. 13, no. Suppl 1, pp. 4–4, 2018.

[90] B. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press, 2016.

[91] J. Noyes, A. Booth, M. Cargo, K. Flemming, R. Garside, K. Hannes, A. Harden, J. Harris, S. Lewin, T. Pantoja, and J. Thomas, "Cochrane Qualitative and Implementation Methods Group guidance series paper 1: introduction," *Journal of Clinical Epidemiology*, vol. 97, pp. 35—38, 2018.

[92] J. L. Harris, A. Booth, M. Cargo, K. Hannes, A. Harden, K. Flemming, R. Garside, T. Pantoja, J. Thomas, and J. Noyes, "Cochrane Qualitative and Implementation Methods Group guidance series paper 2: methods for question formulation, searching, and protocol development for qualitative evidence synthesis," *Journal of Clinical Epidemiology*, vol. 97, pp. 38—48, 2018.

[93] J. Noyes, A. Booth, K. Flemming, R. Garside, A. Harden, S. Lewin, T. Pantoja, K. Hannes, M. Cargo, and J. Thomas, "Cochrane Qualitative and Implementation Methods Group guidance series paper 3: methods for assessing methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings," *Journal of Clinical Epidemiology*, vol. 97, pp. 69—58, 2018.

[94] M. Cargo, J. Harris, T. Pantoja, A. Booth, A. Harden, K. H. and-James Thomas, K. Flemming, R. Garside, and J. Noye, "Cochrane Qualitative and Implementation Methods Group guidance series paper 4: methods for assessingevidence on intervention implementation," *Journal of Clinical Epidemiology*, vol. 97, pp. 59—69, 2018.

[95] K. Flemming, A. Booth, K. Hannesc, M. Cargo, and J. Noyes, "Cochrane Qualitative and Implementation Methods Group guidance series paper 6: reporting guidelines for qualitative, implementation, and process evaluation evidence syntheses," *Journal of Clinical Epidemiology*, vol. 97, pp. 79—85, 2018.

[96] A. Booth, C. Carroll, I. Ilott, L. L. Low, and K. Cooper, "Desperately Seeking Dissonance: Identifying the Disconfirming Case in Qualitative Evidence Synthesis," *Qualitative Health Research*, vol. 23, no. 1, pp. 124–141, 2013.

[97] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp, "Protocol for a systematic literature review of motivation in software engineering," University of Hertfordshire, Tech. Rep. 453, 2006. [Online]. Available: https://uhra.herts.ac.uk/bitstream/handle/2299/992/s73.pdf?sequence=1

[98] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp, "Motivation in software engineering: A systematic literature review," *Information and Software Technology*, vol. 50, no. 9–10, pp. 860–878, 2008.

[99] H. Sharp, N. Baddoo, S. Beecham, T. Hall, and H. Robinson, "Models of motivation in software engineering," *Information and Software Technology*, vol. 51, pp. 219–233, 2009.

[100] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," *Information & Software Technology*, vol. 50, pp. 833–859, 2008.

[101] M. S. Ali, M. A. Babar, L. Chen, and K.-J. Stol, "A systematic review of comparative evidence of aspect-oriented programming," *Information and Software Technology*, vol. 52, no. 9, pp. 871 – 887, 2010.
[102] J. E. Hannay, T. Dybå, E. Arisholm, and D. I. K. Sjøberg, "The effectiveness of pair programming: A meta-analysis," *Information and Software Technology*, vol. 51, no. 7, pp. 1110–1122, 2009.
[103] B. Kitchenham, E. Mendes, and G. Travassos, "Protocol for Systematic Review of Within- and Cross-Company Estimation Models," 2006, version 14. [Online]. Available: https://madeyski.e-informatyka.pl/download/Kitchenham06Protocol.pdf
[104] B. Kitchenham, E. Mendez, and G. H. Travassos, "A systematic review of cross- vs. within-company cost estimation studies," in *EASE 2006*, ser. Evaluation and Assessment in Software Engineering. BCS, 2006.
[105] B. A. Kitchenham, E. Mendes, and G. H.Travassos, "Cross versus within-company cost estimation studies: A systematic review," *IEEE Transactions on Software Engineering*, vol. 33, no. 5, pp. 316–329, 2007.
[106] J. Lavis, G. Permanand, A. Oxman, S. Lewin, and A. Fredheim, "SUPPORT tools for evidence-informed health policy making (STP) 13: Preparing and using Policy Briefs to support evidence-Informed policy-making ," *Health Research Policy & Systems*, vol. 7, 2009.
[107] B. Kitchenham, L. Madeyski, and D. Budgen, "How should software engineering secondary studies include grey material?" *IEEE Transactions on Software Engineering*, 2022. [Online]. Available: https://doi.org/10.1109/TSE.2022.3165938
[108] L. Madeyski and B. Kitchenham, "Would Wider Adoption of Reproducible Research be Beneficial for Empirical Software Engineering Research?" *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 2, pp. 1509–1521, 2017. [Online]. Available: https://doi.org/10.3233/JIFS-169146

Barbara Kitchenham is an Emeritus Professor in the School of Computing and Mathematics at Keele University in the UK. She has worked in software engineering for over 40 years both in industry and academia. She has published over 150 software engineering journal and conference papers. Her main research interest is software measurement and experimentation in the context of project management, quality control, risk management, and evaluation of software technologies. Her most recent research has focused on the application of evidence-based practice to software engineering. In 2019, she was awarded the IEEE Technical Committee Distinguished Women in Science & Engineering (WISE) Leadership Award.

Lech Madeyski is an Associate Professor & Deputy Head of the Department of Applied Informatics at Wroclaw University of Science and Technology, Poland. He has been a Visiting Researcher at Keele University (UK), Brunel University London (UK), and a Visiting Professor at Blekinge Institute of Technology (Sweden). His research focus is on empirical software engineering, data science in software engineering, reproducible research, robust statistical methods. He is a co-founder of *e-Informatica Software Engineering Journal* & International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE). He has published in prestigious journals including, e.g., *IEEE Transactions on Software Engineering*, *Empirical Software Engineering*, *Information and Software Technology*, *Statistics in Medicine*. He is an author of a book "Test-Driven Development: An Empirical Evaluation of Agile Practice" incl. meta-analysis of experiments.

David Budgen is an Emeritus Professor of Software Engineering in the School of Engineering & Computing Sciences at Durham University in the UK. His research interests include software design, design environments, healthcare computing and evidence-based software engineering. He was awarded a BSc(Hons) in Physics and a PhD in Theoretical Physics from Durham University, following which he worked as a research scientist for the Admiralty and then held academic positions at Stirling University and Keele University before moving to his present post at Durham University in 2005. He is a member of the IEEE Computer Society, the ACM and the Institution of Engineering & Technology (IET), and is a Chartered Engineer (CEng).