# Supplementary Material for *SEGRESS: Software Engineering Guidelines for REporting Secondary Studies*

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

**Abstract**—This document includes extended explanations and examples to illustrate how to address all the items in the SEGRESS (Software Engineering Guidelines for REporting Secondary Studies).

**Index Terms**—software engineering, secondary studies, reporting guidelines, PRISMA 2000

## 1 Introduction

In this document we provide further explanations and examples of SEGRESS items. The additional explanations of the SEGRESS items are based on the PRISMA 2020 statement [1], the PRISMA-ScR guidelines [2], the PRISMA-S checklist used to assist the search process [3], the guidelines for qualitative reviews [4], and the guidelines for realist syntheses [5]. Where necessary, we have revised the explanations to fit better with software engineering research.

We have identified examples from a number of software engineering sources. Where possible, we have used examples based on our own studies, and have amended them, if necessary, to show how they could have been better reported by using the SEGRESS guidelines. These systematic reviews are as follows:

1) A quantitative SR assessing the comparative accuracy of single company and cross-company effort estimation models. This is reported in a conference paper [6] and in an extended journal paper [7] and has a protocol available [8]. We use this SR as a basis of a running example of issues associated with risk of individual study bias, risk of missing data and certainty of evidence. We revise the 2007 report so that it fits better with the SEGRESS guidelines and illustrate how the various guideline items work together. We also include some examples from this SR that are not part of the running example. They are quoted as-is from the original SR.
2) A systematic review investigating SR process methods suggested for use by software engineering researchers [9] which has a protocol available [10].
3) A study of meta-analysis methods used in studies reporting families of experiments [11] which has a protocol available.
4) A tertiary study of SE systematic reviews [12] which has a protocol available [13].
5) An extended tertiary study of SE systematic reviews [14] which has a protocol available [15].
6) A review of code smell detection studies [16]
7) A study of the extent of reproducibility found in code smell detection studies [17].

When we have used examples from other researchers, these have been chosen from studies that have impressed us with their rigour and importance. In particular, we rely heavily on two systematic reviews:

1) For qualitative synthesis, we use the systematic review of software engineering motivation conducted by Beecham, Sharp and their colleagues (see [18] and [19]), the planning of which is recorded in their protocol [20].
2) For quantitative analysis, we use the meta-analysis of pair-programming undertaken by Hannay et al. [21]

Other systematic reviews that we have used less frequently, but that are no less important, are (in order of citation):

1) A study of the use of meta-ethnography for qualitative review [22].
2) A systematic review of agile software development, as discussed in two papers [23] and [24].
3) A quantitative meta-analysis of studies comparing fault prediction models [25].
4) A quantitative meta-analysis of studies comparing inspection methods [26].
5) A quantitative study of cost estimation accuracy comparing expert opinion and formal models [27].
6) A review of comparative evidence of aspect-oriented programming [28].
7) A tertiary study updating a previous tertiary study of SE secondary studies [29].

We have concentrated on quantitative and qualitative SRs as sources for our examples, because mapping studies are usually much simpler to report than full SRs.

Many of the quotations we use as examples were obtained from unpublished SR protocols, while quotations

- *B. Kitchenham is with the School of Computing and Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK.*
- *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*
- *David Budgen is an Emeritus Professor in the Department of Computer Science, Durham University, Durham, DH1 4LA, UK.*

*Manuscript received February XX, 2021*

from published papers are relatively short. This was done to avoid copyright infringements; we encourage readers interested in seeing the full discussions to consult the published papers. Readers should note that references to tables and figures in the running example refer to the tables and figures reported in this document, but citations refer to the original document. In contrast, citations and references to tables and figures in quotations relate to the original source document.

The document is organized as follows:

1) Section 2 explains items related to the introduction of a secondary study, i.e., the Title, Abstract, Objectives and Rationale.
2) Section 3 explains items related to reporting the secondary study methods.
3) Section 4 explains items related to reporting the results of secondary studies.
4) Section 5 explains items related to the discussion of results and the limitations of the secondary study.
5) Section 6 explains items related to scientific ethics.
6) Section 7 reports our reflections on producing this document, particularly our experiences of constructing the running example.

## 2 Systematic Review Introduction Items

This section discusses the first three items in a systematic review. It aims to provide an overview information that allows potential readers to find an SR and decide whether they need to read it.

### 2.1 Item 1 Title

PRISMA-2020 Definition: Authors should identify the report as a systematic review.

*Explanation*

Inclusion of "systematic review" in the title facilitates identification by potential users (e.g., researchers, practitioners, managers) and distinguishes the article from a non-systematic literature review.

The title should identify the main objective or research question addressed by the article and should identify whether the article:

- is a mapping study rather than a systematic review (SR),
- is a qualitative review,
- is a tertiary study,
- is an update to an existing SR,
- includes a meta-analysis.

*Examples*

- Cross versus Within-Company Cost Estimation Studies: A Systematic Review [7].
- Motivation in Software Engineering: A systematic literature review [18].
- Reducing test effort: A systematic mapping study on existing approaches [30]
- The effectiveness of pair programming: A meta-analysis [21].

### 2.2 Item 2: Abstract

PRISMA 2020 Definition: See the PRISMA 2020 for Abstracts checklist.

*Explanation*

An abstract should help readers decide whether to access the full report. This means providing information about the main objective(s) or question(s) that the review addresses, methods, results, and implications of the findings. For some readers, the abstract may be all that they have access to. Therefore, results should be presented for all of the main outcomes for the main review's objective(s) or question(s) regardless of the statistical significance, magnitude, or direction of effect. The terms presented in the abstract will be used to index the systematic review in bibliographic databases. Therefore, reporting keywords that accurately describe the review question is recommended.

Table 1 presents the SEGRESS abstract items checklist which has been adapted from the PRISMA 2020 abstract checklist, see [31, Table 2] and [1, Box 2]. SE journals commonly have more stringent limits upon the length of their abstracts than is normal for medical journals. Therefore, we recommend that SE researchers mention only critical issues in their abstracts. Also, medical journals require the abstract to report sources of funding and any personal conflicts of interest, together with the systematic review registration number. Currently, SE journals require any sources of funding to be specified in the Acknowledgments section and there is no registration system for SE SRs, so we have omitted items 11 and 12 in the PRISMA 2020 checklist from our abstract checklist.

*Example*

The following abstract provides a starting point for our running example based on [7]. It based on the revised results reported in the running example in this paper *not* the results reported in the original study:

> *Background: It is important for software companies to know whether or not it is reasonable to use cross-company estimation models to improve the accuracy of their cost estimation process. Aim: This systematic review aims to assess whether cross-company software project data sets can be used to produce project effort estimation models suitable for single companies. Method: We include studies that compare the prediction accuracy of models for a single company based on cross-company data or with prediction based on the single company's own data. We excluded studies where projects were only collected from a small number of different sources (e.g., 2 or 3 companies) or studies where models derived from a single company dataset were compared with predictions from a general cost estimation model. We used keyword searches on six digital libraries (INSPEC, EI Compendex, Science Direct, Web of Science, IEEExplore, ACM Digital Library) to identify papers studying software cost estimation models based on project data from cross-company datasets. The search covered the time period 1990 to November 2006. From 1334 papers, we identified 10 studies. After removing duplicated results (where different studies used the same single company data) and studies that did not report an appropriate effect size, we selected seven studies that reported 12 unique comparisons and reported the percentage magnitude median relative error (MdMRE) to measure accuracy. We calculated the difference between the best MdMRE for the model derived from cross-company data and the best MdMRE model derived from the single*

TABLE 1: Items in the SEGRESS Abstract Checklist adapted from the PRISMA 2020 Abstract Checklist

| Id | Abstract item |
|---|---|
| 1 | Identify the review type. |
| 2 | Provide an explicit statement of the main objective(s) or question(s) the review addresses. |
| 3 | Specify the critical inclusion and exclusion criteria for the review. |
| 4 | Specify the main information sources (such as databases, registers) used to identify studies and the date when each was last searched. |
| 5 | Specify the methods used to assess risk of bias in the included studies. |
| 6 | Specify the methods used to present and synthesise results. |
| 7 | Give the total number of included studies and participants and summarise relevant characteristics of studies. |
| 8 | For quantitative SRs: Present results for the main outcomes, preferably indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (that is, which group is favoured). For qualitative reviews, identify the main scope and themes of any qualitative model(s). For mapping studies and tertiary, summarize the main findings. |
| 9 | For quantitative and qualitative reviews, provide a brief summary of the limitations of the evidence included in the review (such as study risk of bias due to methodological weakness and other issues related to the specific type of review (see GRADE [32] and GRADE-CERQUal [33] respectively). Usually unnecessary for mapping studies and tertiary studies. |
| 10 | Provide a general interpretation of the results and important implications for research and practice. |

> *company data. We used the binomial test to investigate whether the probability of a negative difference (favouring the cross-company model, of which there were two) and the probability of positive difference (favouring the single company model, of which there were ten) were significantly different. Results: The null hypothesis was rejected ($p = 0.039$), but the actual MdMRE difference was not large (median MdMRE difference=6.5 in favour of models obtained from the single company data). The quality of the evidence using the GRADE criteria was assessed as Very Low. There was no evidence that large single company data sets provide better accuracy predictions than small data sets (GRADE-CERQual assessment was Moderate). Conclusions: Companies with only limited data for cost estimation should be cautious of basing estimates on cross-company models. With a data set of 10 or more projects, companies are likely to obtain better estimates from models based on their own data.*

Example 1: Abstract—Running Example Start

We have omitted reporting the number of dataset items because the number of projects do not add up in the same way as the number of patients in a medical trial. Even so, at 400 words the abstract is too long for most SE journals. If reducing length is imperative, we would suggest omitting the specific names of the digital libraries and the eligibility criteria, and reporting only the main result.

### 2.3 Item 3 Rationale

PRISMA 2020 Definition: Authors should describe the rationale for the review in the context of existing knowledge.

#### *Explanation*

Readers should be able to understand why the review was conducted and what the findings from the review might add to existing knowledge.

Authors should:

- Describe the current state of knowledge and its uncertainties.
- Articulate why it is important to perform the review.
- If other systematic reviews addressing the same (or a largely similar) question are available, explain why the current review was considered necessary (for example, previous reviews are out of date or have discordant results; new review methods are available to address the review question; existing reviews are methodologically flawed; or the current review was commissioned to inform a guideline or policy for a particular organisation). If the review is an update or replication of a particular systematic review, indicate this and cite the previous review.
- If the review examines the effects of interventions, also briefly describe how the intervention(s) examined might work.

#### *Example*

In describing the rationale for their SR on cross-company versus single company cost models, Kitchenham et al. [7] first explained why cross-company cost estimation models are useful to industry because of the time needed to assemble a large single company data set. They then described some early studies of the issue that had suggested cross-company costs models were as good as single company cost models (also referred to as within-company cost models). They also pointed out that some subsequent studies were less encouraging, and supported the need for a systematic review as follows:

> *Given the importance of knowing whether or not it is reasonable to use cross-company estimation models to predict effort for single company projects, we conducted a systematic review in order to determine factors that influence the outcome of studies comparing within-company and cross-company models.*

Quote 1: Rationale for Cost Estimation SR [7, pp. 316–317]

### 2.4 Item 4 Objectives

PRISMA 2020 Definition: Authors should provide an explicit statement of the objective(s) or question(s) the review addresses.

#### *Explanation*

An explicit and concise statement of the review objective(s) or question(s) will help readers understand the scope of the review and assess whether the methods used in the review (such as eligibility criteria, search methods, data items, and the comparisons used in the synthesis) adequately address the objective(s). In software engineering, such statements are usually written in the form of questions (“what are the

effects of. . . ?"), but they may be written as objectives ("the objectives of the review were to examine the effects of. . . ").

*Examples*

In their protocol for their SR comparing cross-company and single company effort predictions, Kitchenham et al. [8] identify that previous research results have reported conflicting results and specify the aims of their SR as follows:

> *... the aim of this systematic review is to assist software companies with small data sets decide whether or not to use an estimation model obtained from a benchmarking dataset.*
>
> Quote 2: Cost Estimation SR Aims [8, p. 1]

They then specified four detailed research questions based on those objectives:

> *In order to determine factors that influence the outcome of studies comparing within and between company models, our primary research questions are:*
>
> - *Question 1: What evidence is there that cross-company estimation models are not significantly worse than within-company estimation models for predicting effort for software/Web projects?*
> - *Question 2: Do the characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?*
>
> *Since some studies also compared prediction accuracy between prediction techniques and all the studies used different experimental procedures, we also had two secondary research questions:*
>
> - *Question 3: Which estimation method(s) were best for constructing cross-company effort estimation models?*
> - *Question 4: Which experimental procedure is most appropriate for studies comparing within- and cross-company estimation models?*
>
> Quote 3: Cost Estimation SR Research Questions [8, p. 2]

In their journal paper [7, p. 317], they concatenated research questions 3 and 4 into one research question related to improving procedures for SE cost estimation empirical research.

In the protocol for their study of motivation, Beecham et al. [20] report the development of their research questions as follows:

> *We considered whether our general research question "Does Software Engineer motivation affect software productivity?" is suitable for investigation by systematic review.* Prima facia *this question does not closely match the type suggested by Kitchenham (2004) where the emphasis is on assessing how technology is adopted in/affects software engineering. Our work perhaps relates more closely to the root of the guidelines provided by the medical literature. We can adapt a medical theme,"Assessing the economic value of an intervention or procedure", to "Assessing the economic value of applying motivation approaches in software engineering".*
>
> *Initial research shows very little work in the area of the economics of motivation in software engineering. However, before answering our research question "Does Software Engineer motivation affect software productivity?" we need to know the characteristics of a Software Engineer. This is because we need to understand where Software Engineers are placed in terms of the generic models of motivation found in the psychology, sociology and organisational behaviour texts. When we have a grasp of these characteristics, we can ask: what motivates software engineers; how existing motivation theories are adopted in practice; and how motivation impacts productivity. To ensure that we do not exclude relevant work in this area, we also look at software engineer de-motivators.*
>
> Quote 4: Deriving Reporting Research Objectives for Motivation SR [20, p. 2]

Beecham et al. then defined their research questions as follows:

> *RQ1: What are the characteristics of Software Engineers?*
> *RQ2: What (de)motivates Software Engineers to be more (less) productive?*
> *RQ3: What are the external signs or outcomes of (de)motivated Software Engineers?*
> *RQ4: What aspects of Software Engineering (de)motivate Software Engineers?*
> *RQ5: What models of motivation exist in Software Engineering?*
>
> Quote 5: Motivation SR Research Questions [20, p. 3]

## 3 Specifying the SR Methods

It is much easier to report the methods used for a secondary study if researchers prepare and validate a protocol for their study that reports the development of eligibility criteria and search strings, defines required data items, and reports any trials of the methods used for data extraction and data analysis. Extended discussion of protocol development is beyond the scope of this document, but Shamseer et al. [34] provide guidelines for protocol development consistent with the original PRISMA statement [35].

### 3.1 Item 5 Eligibility Criteria

PRISMA 2020 Definition: Authors should specify the inclusion and exclusion criteria for the review and how studies were grouped for synthesis.

*Explanation*

The criteria used to decide what evidence was eligible or ineligible should be specified in sufficient detail to enable readers to understand the scope of the review and verify inclusion decisions. In particular, ensure that any restrictions with respect to study type or language or publication date are reported and explained in terms of the study objectives.

For qualitative reviews, report any special eligibility requirements used to limit the number of otherwise eligible primary studies. Concepts such as theoretical sampling and purposeful sampling may be applied to the full set of primary studies to restrict the studies included in the synthesis to a manageable set, appropriate for the selected qualitative analysis method.

*Example*

In their protocol for a study of the systematic view process research, Kitchenham and Brereton [10] reported their eligibility criteria in terms of their aims and research questions as follows:

> *The aim of this systematic review is to identify and classify papers related to SLR methodological issues in the context of software engineering, including papers related to quality assessment of primary studies. The inclusion criteria are therefore:*
>
> 1) *That the main objective of the paper which may be a primary, secondary or tertiary study is either to discuss or investigate a methodological issue related to systematic literature reviews, or to discuss or investigate the construction and/or evaluation of quality instruments used to assess primary studies or the general strength of evidence.*
> 2) *Studies discussing how SLRs can or should support EBSE .*
> 3) *That the study must have a software engineering context.*
> 4) *That the study must be published in English.*
>
> *Note that different papers related to the same study will be kept in the set of selected papers but identified as linked papers.*
>
> *The exclusion criteria are:*
>
> 1) *Secondary or tertiary studies whose main objective is to report the results of a systematic review or mapping study. Thus we exclude papers that comment on problems with searches or other processes as part of reporting an SLR or mapping study.*
> 2) *Papers discussing EBSE principles.*
> 3) *Methodological studies with general (i.e. non-software engineering) context.*
> 4) *Papers for which only PowerPoint presentations or extended abstracts are available.*
> 5) *Papers producing guidelines for performing or reporting primary studies as opposed to guidelines for quality evaluation of primary studies.*
>
> *In particular, our selected papers will exclude:*
>
> 1) *The three tertiary studies which were aimed at classifying software engineering SLRs, i.e. Kitchenham et al (2009). Kitchenham et al. (2010), da Silva et al. (2011). These studies discuss the quality of primary studies but are not primarily about the SLR methodology, although they will be referenced in the discussion of related research.*
> 2) *Papers that describe guidelines for SLRs in software engineering (Kitchenham, 2004; Biolchini, 2005. Kitchenham and Charters, 2007). The most recent set of guidelines will be assessed in the light of recommendations obtained from the primary studies in terms of how it should be amended or extended.*
> 3) *Papers reporting studies that developed or evaluated guidelines for performing empirical studies or reporting empirical studies rather than evaluating the quality of empirical studies. For example, the paper by Verner et al. (2009) produced guidelines for performing cases studies, and would be excluded. Similarly, the guidelines for reporting experiments produces by Jedlitschka et al. (2009 are also excluded. In contrast, although their main purpose was to produce guidelines for conducting and reporting of case studies, the paper by Runeson and Höst (2009) includes a checklist for readers which can be considered to be a quality checklist, so we include their paper in our set of included papers.*

Quote 6: Eligibility Criteria [10, p. 8]

In their published report [9], they extend their discussion to report the rationale for each inclusion and exclusion criterion.

### 3.2 Search Process and Strategy: PRISMA-S and Its Relationship with PRISMA 2020 and SEGRESS

Before discussing SEGRESS items 6 and 7 in detail, we introduce PRISMA-S, which is the PRISMA standard for reporting the search and selection process and compare it with the PRISMA 2020 advice.

PRISMA-S was developed by Rethlefsen et al. [3]. It is a checklist of 16 items to support reporting the search process employed in an SR and its search results. PRISMA-S was published prior to the publication of PRISMA 2020, but Rethlefsen et al. report that they communicated with the developers of PRISMA 2020 to keep their checklist in step with the proposed changes to the original PRISMA statement [35].

We report the PRISMA-S checklist in Table 2 and relate it to the relevant PRISMA 2020 items. If PRISMA 2020 does not mention the items we put "NI" in the PRISMA 2020 column.

In addition, PRISMA 2020 identifies some reporting requirements that were not explicitly mentioned in PRISMA-S:

- Item 6 Reference Lists: If reference lists were examined, specify the types of references examined (such as references cited in study reports included in the systematic review, or references cited in systematic review reports on the same, or a similar, topic).
- Item 6 Journals or Proceedings: If journals or conference proceedings were consulted, specify the names of each source, the dates covered, and how they were searched (such as manual searching or browsing online).
- Item 7 Tool support: If natural language processing or text frequency analysis tools were used to identify or refine keywords, synonyms, or subject indexing terms used in the search strategy, specify the tool(s) used.
- Item 7 Validation: If the search strategy was validated—for example, by evaluating whether it could identify a set of clearly eligible studies—report the validation process used and specify which studies were included in the validation set.
- Item 7 Search string development: If the search strategy structure adopted was *not* based on a PICO-style approach, describe the final conceptual structure and any explorations that were undertaken to achieve it (for example, use of a multi-faceted approach that uses a series of searches, with different combinations of concepts, to capture a complex research question, or use of a variety of different search approaches when a specific concept is difficult to define).

The PRISMA-S item "Study Register" is not currently used in a software engineering context. In health-related disciplines, empirical studies, particularly field trials, are expected to be registered. This practice has not been adopted in software engineering, so we have omitted the item from our checklist.

For SEGRESS, we have integrated the PRISMA-S checklist and the PRISMA 2020 Essential elements list and constructed separate checklists for SEGRESS item 6 and item 7. We made the following amendments to PRISMA-S items:

- For consistency with other SEGRESS items, search restrictions should be considered to be part of the eligibility criteria. Our rationale is that limitations on the search are forms of exclusion criteria and it is sensible to report all exclusion criteria in the same place.
- We have generalised the item that Rethlefsen et al. refer to as *Peer Review* to the item *Search Validation*

TABLE 2: Items in the PRISMA-S Search Process Checklist and their Relationship to PRISMA 2020

| PRISMA-S | Item Name | Explanation | PRISMA 2020 |
|---|---|---|---|
| | **INFORMATION SOURCES AND METHODS** | | |
| 1 | Database Name | Name each individual database searched stating the platform for each. | 6 |
| 2 | Multi-data base searching | If databases were searched simultaneously on a single platform, state the name of the platform listing all databases searched. Note. PRISMA 2020 refers to this as a tool to automatically translate search strings used for one database to another. | 7 |
| 3 | Study Register | List any study Registers used | 6 |
| 4 | Online resources and browsing | Describe any online or print source purposefully searched or browsed (e.g., tables of contents, print conference proceedings, web sites) and how this was done. | 6 |
| 5 | Citation Searching | Indicate whether the cited references or citing references were examined for locating cited/citing references (e.g., browsing references lists, using a citation index, setting up email alerts for references citing included studies). | 6 |
| 6 | Contacts | Indicate whether additional studies or data were sought from authors, experts, manufacturers, others. | 6 |
| 7 | Other methods | Describe any additional information sources or search methods used. | NI |
| | **SEARCH STRATEGIES** | | |
| 8 | Full search strategies | Include the search strategies for each database and information source, copied and pasted exactly as run. | 7 |
| 9 | Limits and restrictions | Specify that no limits were used, or describe any limits or restrictions applied to a search (e.g., date or time period, language, study design) and provide justification for their use. | 7 |
| 10 | Search Filters | Indicate whether published search filters were used (as originally designed or modified) and provide justification for their use. | 7 |
| 11 | Prior work | Indicate when search strategies from other literature reviews were adapted or reused for a substantive part or all of the search, citing the previous review(s). | |
| 12 | Updates | Report the methods used to update the search(es) while conducting the SR (e.g., re-running searches, email alerts). | NI |
| 13 | Dates of searches | For each search strategy, provide the date when the last search occurred. If time period of the search was restricted, report the restriction and its justification in the eligibility criteria. | 6 |
| | **PEER REVIEW** | | |
| 14 | Peer Review | Describe any peer review process used to validate the search strings. | 7 |
| | **MANAGING RECORDS** | | |
| 15 | Managing Duplicates | Describe the processes and any software used to identify duplicate records from multiple database searches and other information sources. | NI |
| 16 | Total records | Document the total number of records identified from each database and other information sources. | NI |

to cover both peer review and the string validation recommended in PRISMA 2020.

- We have removed the item related to *Total Records*. Our rational is that reporting the total numbers of records found in each search is not part of the search method, it is related to reporting the search results.
- We have extended the item relating to managing duplicate records to consider duplicate reports of the same study. Our rationale is that in SE authors often report preliminary results in conference papers followed by more extensive journal papers. Such instances need to be identified and managed to avoid over-counting primary studies and introducing errors into any statistical meta-analysis.

### 3.3 The Relationship between Item 6 and Item 7

The report of the search process as specified in Item 6 (Section 3.4) and Item 7 (Section 3.5) of SEGRESS aims to:

- Make the search process as reproducible as possible.
- Allow the SR readers to judge whether the search process was sufficiently complete and up-to-date to have minimised the risk of publication bias.
- Allow the SR readers to judge whether the search process was appropriate to answer the stated objectives and research questions of the SR.

Item 6 *Information Sources* takes a high-level view of the search process and considers the use of mixed approaches to searching (e.g., manual searches and snowballing) to increase the completeness of the search. In contrast, PRISMA 2020 item 7 *Search Strategy* addresses primarily the construction, specification, and validation of search strings used to construct database queries. In Section 3.2, we provide a general introduction for reporting the search process. In the following sections, we provide an extended checklist for item 6 and item 7 and software engineering examples.

### 3.4 Item 6 Information Sources

PRISMA 2020 Definition: Authors should specify all databases, registers, websites, organizations, reference lists, and other sources searched or consulted to identify studies together with the date each source was last searched or consulted.

#### *Explanation*

SEGRESS item 6 (Information sources) requires authors to define the range of information sources searched to give an assessment of the breadth of the search process. It covers searches that are aimed at finding both formally published studies and grey literature (defined as informally published studies, see [36]).

Consideration of PRISMA 2020 and PRISMA-S leads to the following checklist:

1) For published academic and professional articles, report the name of the digital library or platform searched and the URL of the library or platform.
2) Report the date that the library or platform was last searched.
3) Report any tool used to access multiple databases by constructing database specific variants of a single search string (e.g., the ASH tool [37]).

4) Report any special processes used to keep the set of primary studies up to date, such as adding alert requests to digital libraries or platforms.
5) Report the scope of any citation referencing undertaken (i.e., forward or backward snowballing).
6) Report any specific reference lists searched, such as those produced by related secondary studies.
7) Report any manual searching or browsing of specific journals or conference proceedings.
8) For unpublished material (i.e., grey literature):
   - report the URL of any web sites searched,
   - report any individuals approached and explain why,
   - report the name of any organisations or industry sources approached.
9) Report how duplicate records were identified and handled. Consider both duplicate records found by searching different digital libraries, as well as reports of the same study found in different articles.

If SE adopts the idea of registering primary studies planning to evaluate SE technologies, SR reviewers should report any use of such registers.

*Example*

In their study of meta-analysis in families of experiments, Kitchenham et al. report and justify their choice of information sources as follows:

> *In order to address our research questions, we needed to identify papers that reported the use of meta-analysis to aggregate individual studies, reported the results of the individual studies in detail, and were published in high-quality journals. To achieve our search process strategy, we decided to limit our search for families of experiments to the following five journals:*
>
> - *IEEE Transactions on Software Engineering (TSE).*
> - *Empirical Software Engineering (EMSE).*
> - *Journal of Systems and Software (JSS).*
> - *Information and Software Technology (IST).*
> - *ACM Transactions on Software Engineering Methodology (TOSEM).*
>
> *We restricted ourselves to these journals because they all publish papers on empirical software engineering, and all have relatively high impact factors (among SE journals). These are, therefore, highly respected journals, and we should expect the quality of papers they publish to be correspondingly high.*
>
> Quote 7: Information Sources for Families of Experiments [11, p. 356]

In their protocol, Kitchenham et al. report that all five journals were accessed by means of the Scopus platform, and the search was checked by using the DBLP database.

### 3.5 Item 7 Search Strategy

PRISMA 2020 Definition: Authors should present the full search strategies for all databases, registers and websites including any filters and limits used.

*Explanation*

SEGRESS item 7 (Search Strategy) requires authors to define the details of how the search strings were constructed and validated, and to specify the individual search strings used for each digital library or platform. The rationale is that reporting the full details of all search strings (such as the full, line by line search strategy as run in each database) should enhance the transparency of the systematic review, improve reproducibility, and enable a review to be more easily updated.

Consideration of PRISMA 2020 and PRISMA-S leads to the following checklist:

1) Explain how the search string(s) were developed linking the explanation to the SR goals and eligibility criteria as necessary.
2) Report any search filters used for specific databases.
3) If natural language processing or text frequency tools were used to identify keywords, specify the name and version of the software used, how the software was trained and used, and report any available information concerning its performance and reliability.
4) Explain how the search strings were validated, for example by peer review or by reference to a known set of primary studies.
5) Report the full search string used for each digital library, platform, or tool. Cutting and pasting the string will ensure correctness. Also, specify any filters applied to the set of papers after a broad search, for example restricting Scopus search outcomes to papers from Computer Science journals.

*Example*

In the protocol of their study of software engineer motivation, Beecham et al. [20] discuss the method they used to construct their search strings as follows:

> *The following details of the population, intervention, outcomes, and experimental designs of interest to the review will form the basis for the construction of suitable search terms later in the protocol (Section 3.2). We note however, that not all research questions require intervention.*
>
> *Population: Software Engineers*
> *Intervention: motivation approaches, productivity measures*
> *Outcomes of relevance: Software Engineer characteristics; motivational factors; results of applying motivational methods, change in productivity (to include quality and timescales), models of motivation.*
> *Experimental design: Empirical studies, theoretical studies, expert observation, experience reports.*
>
> Quote 8: Method Used to Construct Search Strings [20, p. 3]

Later, they report their detailed strategy for developing search strings to be:

> a. *derive major terms from the questions by identifying the population, intervention and outcome;*
> b. *identify alternative spellings and synonyms for major terms;*
> c. *check the keywords in any relevant papers we already have;*
> d. *when database allows, use the Boolean OR to incorporate alternative spellings and synonyms;*

> e. *when database allows, use the Boolean AND to link the major terms from population, intervention and outcome.*
>
> Quote 9: Strategy for Developing Search Strings [20, p. 4]

In the Appendix to their protocol, they report the search strings used for each of the eight databases they searched as well as the date searched and search outcomes.

Kitchenham et al. used a similar approach in their comparison of single company and cross-company cost estimation models (see [8] and [6]). This method of developing search strings was suggested in Kitchenham's original SR guidelines [38], which recommended the use of PIO (Population, Intervention, Outcome). This approach, with an additional letter "C" for Contrast/Characteristics [1] to give the acronym PICO, is mentioned in the PRISMA 2020 guidelines.

However, the PIO/PICO approach did not prove as useful in software engineering SRs, as it did in medical SRs. There were several reasons for this:

- The concept of population, intervention, and outcomes is not well-suited to SE studies, which are mainly mapping studies and qualitative reviews.
- Most software engineering digital libraries (Scopus excepted) do not handle Boolean constructs correctly, leading to many false positives when long Boolean strings are employed.

The more common approach to creating search strings for SE systematic reviews involves:

1) Restricting the searches either to SE and CS libraries such as the ACM library and IEEExplore or to CS related articles in general digital libraries such as Scopus.
2) Using the main keywords based on the topic area (which could be systematic reviews and mapping studies for some tertiary studies) with very simple Boolean strings.
3) Including start date limits if the concept can be tracked to a specific year.
4) Using forward and backward snowballing to increase coverage.

As an example of the approach, see Section 2.3.2 in a SR by Lewowski and Madeyski [16].

This approach is acceptable in SEGRESS assessment, as is the classic PICO/PIO approach.

### 3.6 Item 8 Selection Process

PRISMA 2020 Definition: Authors should 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.

#### *Explanation*

Study selection is typically a multi-stage process in which potentially eligible studies are first identified by screening titles and abstracts, then assessed through full text review and, where necessary, by contact with the authors of the candidate primary study. Increasingly, a mix of screening approaches might be applied (such as automation to eliminate records before screening or prioritise records during screening).

Authors should describe in detail the process for deciding how records retrieved by the search were considered for inclusion in the review, to enable readers to assess the potential for errors or bias in the selection process.

Authors should report:

1) How many reviewers screened each primary study (title/abstract) retrieved, whether multiple reviewers worked independently (that is, were unaware of each other's decisions) at each stage of screening or not (for example, records screened by one reviewer and exclusions verified by another), and any processes used to resolve disagreements between screeners (for example, referral to a third reviewer or by consensus).
2) Any procedures used to check or assess the consistency of screeners.
3) Any processes used to obtain or confirm relevant information from the authors of a specific primary study.
4) If abstracts or articles required translation into another language to determine their eligibility, report how these were translated (for example, by asking a native speaker or by using software programs).
5) The details of any software tools used as part of the selection process (including versions where appropriate). For machine intelligence-based tools report how the tool was trained and its accuracy statistics.

#### *Examples*

For qualitative reviews, the process of selection may be integrated with either the data extraction or quality extraction process. For example, da Silva et al. [22] discuss the use of meta-ethnography to synthesize primary studies. From a set of five studies that addressed their question of interest, they based their final selection on the quality of the studies (based on the quality assessment tool reported in [23]) and whether the remaining set of selected studies were appropriate for meta-ethnography, saying:

> *We decided to exclude TP5 because of its low score in the quality assessment. After removing TP5, considering the similarities of the four remaining studies, we concluded that the studies formed a coherent set adequate for a meta-ethnography.*
>
> Quote 10: Selection Method for Qualitative Review used by da Silva et al. [22, p. 155]

However, other qualitative reviews use fairly standard selection methods. For example, Dybå and Dingsøyr [23] report their selection process as a four-stage process defined as:

1) Stage 1: Identify candidate studies by searching 8 SE databases and 3 relevant conference proceedings.
2) Stage 2: Exclude studies on the basis of titles.
3) Stage 3: Exclude studies on the basis of abstracts.
4) Stage 4: Obtain full text of studies and critically appraise them, excluding lessons-learned and single-practice studies.

1. Different sources define PICO differently.

They reported using EndNote and Excel sheets to manage the citations in Stage 1. In stage 2 they reported their process to be based on the two authors working together to go through all the study titles. At Stage 3, their process involved a third researcher and each study abstract was assessed by two researchers. At that stage, they calculated the Kappa statistic to measure agreement and then resolved any disagreements by discussion among all three researchers. The only unusual aspect of their process compared with the process usually used by quantitative studies was that for Stage 4, they integrated the exclusion of lessons-learned and single-practice studies with the evaluation of study quality.

### 3.7 Item 9 Data Collection Process

PRISMA 2020 Definition: Authors should specify the methods 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.

#### *Explanation*

The methods used for data collection should be reported in sufficient detail for the process to be reproducible and for readers to be able to assess the risk of error or bias in collected data.

Authors should report:

1) How many reviewers collected data from each report, whether multiple reviewers worked independently or not, and any process used to resolve disagreements.
2) Any reviewer agreement statistics calculated.
3) If any primary study authors were contacted about missing data, what data they were asked to provide and whether the required information was obtained.
4) Any software tool that was use to support data collection, how the tool was used and how it was trained (if training was required). In addition, the risks associated with using the tool should be discussed.
5) How any required primary study translation was performed.
6) How data from multiple reports about the same primary study were integrated.

#### *Examples*

In their report of a study of SR process research, Kitchenham and Brereton [9] reported that Kitchenham extracted basic citation information for each paper, while both authors extracted primary study specific data for each paper that was based on a preliminary categorization of the known studies. They used an Excel spreadsheet that was trialled by both authors as part of the protocol development. However, they needed two different types of form, for two different types of study:

1) For papers that covered a specific process improvement topic and included limited outcomes and recommendations (such as the use of textual analysis tools to aid primary study selection, and a pseudo gold-standard to assist the search process), the original form was used for each study.
2) For discussion papers and lessons-learned papers that had a broad scope, a special text-based extraction form was set up to allow individual textual elements to be extracted.

Kitchenham and Brereton also report that both authors extracted data from topic-specific papers independently, and discussed disagreements until agreement was reached. For textual data extraction from discussion and survey papers, Kitchenham performed the extraction which Brereton subsequently checked. The Kappa statistic was used to check data extraction agreement.

### 3.8 Item 10a Outcome Data

PRISMA 2020 Definition: List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain 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.

The item applies to quantitative and qualitative SRs but not mapping studies (which are addressed by item 10b) because mapping studies do not analyse outcome measures, they characterise primary studies.

#### *Explanation*

Authors should define all the data (numerical or textual) that needs to be collected from each primary study to answer all the research questions related to the findings of the primary studies.

In quantitative experiments, authors of primary studies may report various outcome metrics related to software task effort, task elapsed time and measures of task correctness. Different basic outcomes may be measured in different ways in different studies. In qualitative primary studies, primary study authors may develop conceptual models based on abstract concepts or themes and their relationships. These will be reported in terms of study-specific labels and definitions. In both cases, review synthesis may be subject to error if differences in definitions between different primary studies are not properly reported.

Review authors need to ensure that data definitions are sufficient to ensure data extraction is reliable and that other researchers can use the definitions in other studies, and can replicate or update the review. Readers need to understand the collected data well enough to confirm that the subsequent analysis or synthesis respects properties of the collected data. For example, some SE meta-analyses failed to appreciate the need to adjust effect sizes if software experiments employ replicated measures [39].

Authors need to:

1) List and define the outcome-related data that will be collected for each primary study. This should include extracting the definitions of the metrics used in experiments, and both the label of, and the definitions of, any concepts reported in qualitative models as well as any other qualitative outcomes.
2) If any changes were made to the data definitions during the review, specify the changes and the rationale for the changes. For examples, data definitions defined in the protocol may be changed or refined if outcome metrics

reported in primary studies are more complex than expected.

#### *Examples*

In their protocol for their study of SR process research, Kitchenham and Brereton [10] specify the data to be collected from each primary study as follows:

> *Primary study ID Author(s) Title Publication venue Date of publication Publication details for journal (Volume and Issue) Page numbers (if available)*
>
> *The primary study specific data that will be extracted is based on a preliminary categorisation of the known studies. It includes:*
>
> 1) *Type of Paper: Problem identification and/or problem solution (PI) or Experience Paper, Opinion Survey or Discussion paper (E)*
> 2) *Scope of the study: Mapping studies/Conventional Systematic review/Both/Neither (which must be specified)*
> 3) *Summary of aims of Study*
> 4) *Topics covered (NOT mutually exclusive):*
>    a) *Educational issues: Yes/No*
>    b) *SLR Participant Viewpoint: Experience Researcher (E) / Novice (N) /Not specified (NS)*
>    c) *Research questions: Yes/No*
>    d) *Protocol Development: Yes/No*
>    e) *Search processes: Yes/No*
>    f) *Search validation/evaluation: Yes/No*
>    g) *Selection processes: Yes/No*
>    h) *Quality evaluation of primary studies: Yes/No*
>    i) *Data Aggregation: Yes/No*
>    j) *Data Synthesis: Yes/No*
>    k) *Reporting: Yes/No*
> 5) *Method proposed: Name or description (e.g. Quasi-Gold Standard, Visual text Mining)*
> 6) *Validation/Evaluation performed: Yes/No*
> 7) *Actual Validation method (as judged by each researcher): Experiment, Quasi Experiment, Tertiary Study, Case study, Data Mining, Opinion survey (Interview), Opinion Survey (Questionnaire), Lesson Learnt, Example, Other (to be specified)*
> 8) *Claimed Validation method (as specified by authors of paper)*
> 9) *Summary of main results. Note details of lessons learnt and opinion survey results will be collected in a separate word file.*
> 10) *Any process recommendations (suggested by data extractors).*
>
> Quote 11: Reporting Data Items [10, p. 10]

In the above list the outcome data are items 5 to 9.

Continuing our running example based on the comparison of the accuracy of single company and cross-company models:

> *The main outcome measures collected from each study are specified and justified in Table 3.*
>
> Example 2: Output Data Items—Running Example Continued

### 3.9 Item 10b Other Data

PRISMA 2020 Definition: List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources) and describe any assumptions made about any missing or unclear information.

This item is critical for mapping studies since their main purpose is to identify the main characteristics of each primary study and provide frequency plots of the number of primary studies in different categories. However, both quantitative and qualitative reviews frequently make use of context information to help explain their results.

#### *Explanation*

Any other data collected from each primary study also needs to be fully defined. Such data includes categorical variables describing context factors such as types of participant, the setting of the study, the type of software materials and the specific interventions used.

Primary study data that explains the context of a study, is useful for several purposes:

1) Seeking explanations for differences in reported outcomes,
2) Identifying limitations concerning the scope of the evidence and qualifying recommendations

Such data needs to be well-defined to ensure that the review can be replicated and that readers understand the subsequent data analyses and syntheses, and the discussion of the results.

Authors need to:

1) List and define the all context related data items.
2) Mapping studies authors should specify the research questions that the data item addresses.
3) Specify any changes to data definition during the conduct of the review and the rationale for the changes.

#### *Examples*

In Quote 11, items 1 to 4 in the numbered list identify *Other Data* corresponding to contextual information about each primary study. Item 10 is part of the data synthesis activity.

Continuing our running example based on the comparison of the accuracy of single company and cross-company models:

> *The other measures collected from each primary study are specified in Table 4.*
>
> Example 3: Output Data Items—Running Example Continued

### 3.10 Item 11 Study Risk of Bias Assessment

PRISMA 2020 Definition: Authors should 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.

This item is not usually required for mapping studies.

#### *Explanation*

The PRISMA 2020 guidelines [1] and the Cochrane Collaboration Handbook of Review [40] both criticise tools for assessing the limitations of primary studies that extend beyond issues that have the *potential to bias findings* and/or combining individual items to construct a numerical score. Both of these methods were encouraged in the original software engineering SR guidelines (i.e., [38] and [41]) but are now considered unhelpful.

TABLE 3: Outcome Measures

| Measure | Scope | Justification |
|---|---|---|
| Median Magnitude Relative Error (MdMRE) | To be extracted for each single company, for the best fitting effort estimation model obtained from the single company and the best fitting effort estimation model obtained from the cross company data. | Addresses the main research questions by providing the mean to compare the accuracy of the estimations model. |
| p-value of statistical tests | To be extracted for each single company dataset. | To summarize the conclusions drawn by each study. |
| The names of the cost models used in the statistical tests | Collect the name of each cross-company model used in each primary study and the name of the single company model if it was not part of the cross-company dataset. | To investigate any discrepancies and duplications among different primary studies and ensure consistency for primary study synthesis. |

TABLE 4: Contextual Data

| Measure | Scope | Justification |
|---|---|---|
| Size of single company dataset (number of projects) | Collected for each single company investigated in each primary study. | Dataset size is a possible driver of cost estimation model accuracy. |
| Size of the cross-company dataset (number of projects) | Collected for each comparison in each primary study. | Dataset size is a possible driver of cost estimation model accuracy. |
| Dataset name | Collected the cross-company dataset name and the single company name if different. | Required to identify the range of cross-company data sets used and possible duplication among studies. |
| Cost estimation modelling method | Collected for each comparison reported in the study. | Required to identify the range of estimation methods used, possible duplication among studies and to investigate difference between study results. |
| Estimation Validation method | For the estimates derived from both the cross-company data and singe company data, the method of assessing the accuracy. | Different validation regimes can influence accuracy measures. |
| Statistical significance | Identifies Whether there was a significant difference between estimate accuracy. | To summarize the conclusion of each primary study. |
| Maximum and minimum effort | Collected for each study and each cross company and single data set company, the maximum and minimum project effort. | To assess the effort heterogeneity (referred to as *HetEff* for single company data sets, we calculated the range of effort values for the single company projects divided by the range of effort values for the cross-company projects. |

In addition to defining the criteria used to assess risk of bias (RoB), authors need also to report how criteria were extracted to assure the reader that the problem of experimenter biased was minimised.

Authors need to report:

1) The tool(s) (and version) used to assess risk of bias in the included studies. In this context *tools* are usually lists of criteria expressed as questions about the study that may be organised into sub-groups (referred to as domains) related to different RoB issues.
2) The methodological domains/components/items of the risk of bias tool(s) used.
3) Whether an overall risk of bias judgement that summarised across domains/components/ items was made, and if so, what rules were used to reach an overall judgement.
4) Any adaptations to an existing tool to assess risk of bias in studies that were made (such as omitting or modifying items).
5) The content and details of any new tool developed for the specific review.
6) How many reviewers assessed risk of bias in each study, whether multiple reviewers worked independently (or whether other, less rigorous, methods were used such as assessments performed by one reviewer and checked by another), and any processes used to resolve disagreements between assessors.
7) Any processes used to obtain or confirm relevant information from study investigators.
8) How any automation tool used to assist RoB assessment was used (such as machine learning models to extract sentences relevant to risk of bias from articles), how the tool was trained, and details on the tool's performance and internal validation.

### *Examples*

SE researchers often make use of quality assessment rather than risk of bias evaluations in software engineering SRs. However, PRISMA-2020 is clear that the important issue is to assess risk of methodological bias for each primary study. Primary study risk of bias (RoB) provides information that is used as part of certainty assessment, see Section 3.20. Initially RoB was restricted to randomised field experiments, which are referred to as randomised controlled trials (RCTs) in the medical literature. However, in SE, most experiments are laboratory studies, and field studies are mostly qualitative studies such as case studies and ethnological studies, or opinion surveys. In addition, there are many cost estimation and fault prediction studies studies that analyse industry or open source data sets.

Generally, the method for assessing RoB is to identify various *domains* that describe aspects of methodological risk and associate a set of questions with each domain.

Numerical assessments of risk of bias are no longer considered useful because critical or serious risks of bias with respect to methodology cannot be cancelled out by other low-risk issues. The aim is to have an overall assessment of risk of bias for each domain and for the primary study as a whole. One approach is to assess each domain in terms of the options: Very Low RoB, Low RoB, Moderate RoB, High RoB, No Information. Some approaches use a more refined scale, but we suggest using only a four-point scale for consistency with later use of the results when assessing the certainty in the body of evidence. Continuing our running

example, we present a list of items used to assess risk of bias and other risk related criteria

*We used the following questions to assess risk of bias in our primary studies:*

1) *Is the data analysis process appropriate?*
   a) *Was the data investigated to identify outliers and to assess distributional properties before analysis? Yes (Low RoB), Probably Yes (Moderate RoB), No (High RoB)*
   b) *Were statistical tests used to assess the comparative accuracy estimates appropriate: Yes (Low RoB), Some concerns (Assessment of RoB depends on the potential bias of tests used), No Statistical tests applied (High RoB)*
2) *Did studies carry out a sensitivity or residual analysis?*
   a) *Were the resulting estimation models subject to sensitivity or residual analysis? Yes (Low RoB), Probably Yes (Moderate RoB), No (High RoB)*
   b) *Was the result of the sensitivity or residual analysis used appropriately to discuss results or revise the estimation models? Yes (Low RoB), Probably Yes (Moderate RoB) No (High RoB)*
3) *Were accuracy statistics based on raw data? No (Low RoB), Yes (High RoB)*
4) *How good was the comparison method?*
   - ***How was the single company selected***
   - *How was the accuracy of the estimation models assessed? Random subsets (Low RoB), leave-one-out (Moderate RoB), no hold out (High RoB)*
5) ***What were the sizes of single company data set and the cross-company data set? Report the number of studies***

*The questions relating to the validity of the statistical methods used by each primary study are used to assess the risk of primary study bias. Other questions (shown in bold text) are used to help assess the certainty in the body of evidence.*

*For all but two of the primary studies, two of the three authors were assigned to extract the criteria, at random. We performed our assessments independently, and discussed any disagreements until we reached an agreement. The two remaining primary studies were co-authored by two of the SR authors. For these primary studies the third SR author acted as the main assessor. The assessments for these studies were checked by the each of the other study authors, with the assessments of the independent assessor taking precedence if any disagreements could not be resolved.*

Example 4: Risk of Bias Assessment —Running Example Continued

For qualitative studies, Dybå and Dingsøyr [23] developed a checklist for assessing quality based on the Critical Appraisal Skills Programme (CASP) tool for qualitative research. Although there is a more recent version of CASP, the checklist produced by Dybå and Dingsøyr [23] is still useful because it is focused on software engineering studies:

1) Is the paper based on research (or is it merely a "lessons learned" report based on expert opinion)?
2) Is there a clear statement of the aims of the research?
3) Is there an adequate description of the context in which the research was carried out?
4) Was the research design appropriate to address the aims of the research?
5) Was the recruitment strategy appropriate to the aims of the research?
6) Was there a control group with which to compare treatments?
7) Was the data collected in a way that addressed the research issue?
8) Was the data analysis sufficiently rigorous?
9) Has the relationship between researcher and participants been adequately considered?
10) Is there a clear statement of findings?
11) Is the study of value for research or practice?

The first three questions are "screening questions". If the answer to those questions is "No" it is not worth continuing with the other questions. In their Appendix B, Dybå and Dingsøyr [23] provide detailed sub-questions that should be considered when answering each top level question.

We suggest checklist users convert answers to questions 4-11 from simple Yes or No answers to Very Low, Low, Moderate, High RoB, since in many cases the methodology used may not be ideal, but may be less risky than doing nothing. In addition, RoB-based answers allow a simple conversion to an overall RoB assessment.

### 3.11 Item 12 Effect Measures

PRISMA 2020 Definition: Authors should specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.

This item applies only to quantitative meta-analyses and mixed methods analyses. Effect measures specify the value of an outcome measure such as effort, elapsed time, or fault rate, used to assess a software engineering task or process.

The effect measures used in a systematic review should be specified in the protocol and reported in the review. Any changes to the effect sizes during conduct of the review should be specified and justified.

#### *Explanation*

In order to understand study syntheses and results, readers need to know which effect size(s) were used. For an overview of common effect sizes see [42], while for guidelines for effect size magnitude interpretation see [43].

Authors need to:

1) Explain the choice of effect size and justify the use of any non-standard effect size.
2) Specify for each outcome or type of outcome (such as binary, continuous) the effect measure(s) (such as risk ratio, mean difference) used in the synthesis or presentation of results.
3) State any thresholds or ranges used to interpret the size of effect (such as minimally important difference; ranges for no/trivial, small, moderate, and large effects) and the rationale for these thresholds.
4) If synthesised results were re-expressed to a different effect measure, report the methods used to re-express results. For example, Kitchenham et al. [11] found that many meta-analyses calculated standard effect size differences but then converted the metric into the point bi-serial correlation coefficient for aggregation.

#### *Examples*

In their investigation of primary studies reporting the prediction models for fault-prone software modules/components, Shepperd et al. [25] discuss weaknesses of the F measure and ROC curves and say:

> *For this reason, we advocate a binary correlation coefficient variously known as the Matthews correlation coefficient (MCC) or $\phi$. Unlike the F-measure, MCC is based on all four quadrants of the confusion matrix.*
>
> Quote 12: Choice of Effect Size Measures [25, p. 607]

Hannay et al. [21] report their effect size as follows:

> *In this meta-analysis, we used Hedges' g as the standardized measure of effect size. Like Cohen's d and Glass' D, Hedges' g is simply the difference between the outcome means of the treatment groups, but standardized with respect to the pooled standard deviation, $s_p$, and corrected for small sample bias.*
>
> Quote 13: Choice of Effect Size Measures [21, Section 2.5, p. 1112]

Continuing our running example of the comparison of single company and cross-company effort predictions:

> *The main effect size used in this SR is the* difference *between the MdMRE for the estimation model based on the single company project data and the estimation model based on the cross-company project data. MRE is the absolute relative error of the difference between an actual value and an estimate of that value, calculated as follows:*
>
> $$MRE_i = \frac{|(x_i - \hat{x}_i)|}{x_i} \quad (1)$$
>
> *where $x_i$ is the actual value and $\hat{x}_i$ is the estimate. MdMRe is median of values of $MRE_i$ collected for a specific set of estimates. The mean $MRE$ is known to be a biased metric for comparing two different estimation models, but the median is a less biased metric and the goal of this analysis is to evaluate the best fitting model for a specific dataset.*
>
> *Some primary studies used the best cross-company estimation* method *to derive the single company model used in statistical tests, while others used the estimation* method *that produced the best single company model. To investigate these different approaches, for each primary study and single company estimate, we extracted the MdMRE calculated for estimates based on the most accurate cross-company model ($CC_Effect$), the MdMRE based on the single company model used in any statistical tests ($SEffect$), and the MdMRE based on the best single company model ($BestSEffect$). From this data we calculated two difference based effect sizes:*
>
> $$Difference1 = CCEffect - SEffect \quad (2)$$
>
> $$Difference2 = CCEffect - BestSCEffect \quad (3)$$
>
> *In both cases a negative difference suggests that the cross-company estimation model was more accurate than the estimation model derived from the single company data.*
>
> Example 5: Effect Measures—Running Example Continued

### 3.12 Item 13 Analysis and Synthesis Methods General Issues

In PRISMA 2020, this item is intended to address the pre-planned elements of the synthesis of quantitative systematic reviews and mixed-methods reviews. For SEGRESS, we have extended this item to cover qualitative reviews and mapping studies, but for these reviews the individual items have very different definitions and implications. In particular, many of the sub-items are not relevant for mapping studies. Users of SEGRESS should make sure that they adopt the definition of the item appropriate to the type of SR they need to report.

### 3.13 Item 13a Analysis and Synthesis Eligibility

PRISMA 2020 Definition: Authors should describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item 5)).

#### *Explanation*

Before undertaking any statistical or qualitative synthesis or analysis, decisions must be made about which studies are eligible for each planned synthesis. These decisions may involve subjective judgements that could alter the result of a synthesis or analysis. Reporting the selection processes (whether formal or informal) and any supporting information is recommended to ensure transparency of the decisions made in grouping studies for synthesis.

For quantitative systematic reviews, some primary studies may include more outcome variables than others so analysis of some variables may be based on fewer data points than others. Also, if reviewers are concerned only with meta-analysis of trustworthy studies, they may need to exclude primary studies with high risk of bias.

For qualitative reviews, if techniques such as meta-ethnography are being used, it may be essential to restrict the synthesis to a manageable subset of the eligible studies. Researchers using grounded analysis may need to introduce eligible studies into the synthesis activity until they have achieved theoretical saturation (i.e., continued sampling from available primary studies and including primary studies in model building until all concepts in the theory are well-developed). In such cases, reviewers should try to define in advice the procedures they plan to use to organise the primary study selection process e.g., date order, random order, or coverage of subgroups of primary studies.

For mapping studies, primary studies are classified against various categories and then analysed with respect to those categories. Different data may be collected for different categories of primary studies leading to different analyses.

#### *Examples*

In their study of SR process research, Kitchenham and Brereton [9] split their set of primary studies into subsets that addressed similar issues. Studies that addressed the SR processes in general such as lessons learnt, surveys, and discussion papers were separated from studies that addressed a specific issue with the SR process or particular topics such as education. For the studies addressing a specific issue, Kitchenham and Brereton report:

> *Studies covered by the classification scheme were grouped into sets of studies addressing similar issues ... Within each category, papers were grouped with respect to the specific technique being proposed or the particular task in the SR process.*
>
> Quote 14: Definition of Synthesis Categories [9, p. 2055]

Continuing our running example of the comparison of single company and cross-company effort predictions:

> *The primary studies for in our statistical synthesis were based on the following criteria:*
> 1) *We restricted ourselves to studies that used a statistical test to compare the accuracy of cross-company and single company models. This excluded Study 1 and Study 7.*
> 2) *We restricted ourselves to to only one study for a specific cross-company and single company combination. This raised the issue of including Study 2 or Study 6a. We selected study 6a because the the method of validation was more clearly specified.*

Example 6: Primary Study Synthesis Eligibility Criteria—Running Example Continued

### 3.14 Item 13b Analysis and Synthesis Data Preparation

PRISMA 2020 Definition: Authors should describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.

#### *Explanation*

Authors may need to prepare the data collected from studies so that it is suitable for presentation or to be included in an analysis or a synthesis. Data preparation may involve transforming outcome measures into a common metric and/or procedures for handling missing values.

#### *Examples*

In their study of prediction models to identify fault-prone components, Shepperd et al [25] used a method of reconstructing the confusion matrix from performance measures such as recall, precision, F etc. and converting the resulting confusion matrix into the Mathews correlation coefficient (MCC) for analysis. The transformation method is reported in Bowes et al. [44].

### 3.15 Item 13c Graphical and Tabulation Methods for Analysis, Synthesis, and Individual Studies

PRISMA 2020 Definition: Authors should describe any methods used to tabulate or visually display results of individual studies and syntheses.

#### *Explanation*

Presentation of study results using tabulation and visual display is important for transparency (particularly so for qualitative reviews, and quantitative SRs that do not use formal meta-analysis, but investigate outcome patterns in the data).

For quantitative studies, if meta-analysis is adopted, forest plots can be used both to summarise the effect sizes and their variance for each of individual primary studies and the overall summary effect size from the meta-analysis. If meta-analysis is not adopted, outcome data can be summarized along with primary study characteristics in Tables with primary studies grouped according to whether effect sizes are positive, inconclusive or negative. For quantitative reviews, particularly meta-analysis, the types of synthesis and methods of presenting them are well-defined and can be specified in advance (for example, in the protocol) to help reduce the risk of fishing for significant results by multiple sub-group testing[2].

For qualitative studies, qualitative results are often displayed as conceptual models comprising boxes (or bubbles) linked by lines, where the bubble names identify concepts or themes and the lines (which are often directed) represent the relationships among the concepts. In some qualitative primary studies the outcomes are lists of themes with their definitions, which are often accompanied with a count of the frequency with which the themes were encountered. However, pre-specification of specific graphical representations is not always possible for qualitative reviews, where themes and relationships cannot always be identified in advance. In the Methods section, the basic representation approaches should be defined and the details of the specific representation can be reported when the results are presented and discussed.

For mapping studies, the main results are the categories assigned to the primary studies. These are sometimes displayed as simple tables of categories and their values for each primary study. In addition, they are sometimes presented as bubble plots that display the frequency with which primary studies display the same values of three different categorical values, based on one common categorical y-value and two different categorical x-values [45]. In most cases, the details of the graphical or tabulation representations are best discussed at the point when the results are reported, since they should be related directly to the information needed to address a specific question and there is no concern about multiple statistical tests for mapping studies. In the Methods section, it is usually sufficient to say that the data addressing each research question will be displayed in tables, box plots, or graphs as appropriate. It is only necessary to discuss novel display methods in the Methods section.

Authors should:

1) Report chosen tabular structure(s) and graphs used to display results of individual studies, analyses and syntheses, along with details of the data presented.
2) Define and justify any groupings used to order the presented data.
3) Justify the use of any non-standard graphs.

#### *Examples*

Although they did not specify their choice of graphical representation in the Methods section, Hannay et al. [21] explained their graphical display as follows:

> *Fig. 1 shows Forest plots of the standardized effects for each of the three outcome constructs. The studies are sorted according to the relative weight that a study's effect size receives in the meta-analysis... The rightmost columns in Fig. 1 show these weights according to the fixed-effects and random-effects models...*

Quote 15: Choice of Graphical Representation [21, Section 3.2, p. 1114]

2. This issue applies to meta-analyses just as it applies to individual experiments.

Forrest plots illustrate well the benefits of meta-analyses, as the width of the 95% confidence interval is generally much narrower than in the case of individual studies, see, e.g., Chapter 9 in [46].

### 3.16 Item 13d Methods Used for Analysis and Synthesis

PRISMA 2020 Definition: Authors should describe any methods used to synthesise 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 statistical heterogeneity, and software package(s) used.

#### *Explanation*

Readers need to be informed about which methods were used to analyse and synthesise data from primary studies, in order to understand the analysis and synthesis results. Details and justification of choices made about the analysis procedures need to be specified to support SR reproducibility.

For quantitative systematic reviews, it is important that the tests used for the main hypotheses are planned in advance and reported in the protocol. This is to minimise the problem of fishing for significant results. Researchers may, of course, undertake additional analyses to respond to the specific data, but such tests need to be specified as being additional unplanned tests. Regardless of the chosen synthesis method(s), authors should provide enough detail for readers to be able to assess the appropriateness of the selected methods and to reproduce the reported results (with access to the data).

For mapping studies, there is no necessity for pre-specifying analyses, although reviewers need to ensure that the data planned for extraction addresses the research questions. In the methods section of the SR report, researchers need to identify which research questions will be addressed by which data items. More detailed information should be supplied only if complex or unusual analyses are required.

For qualitative reviews, qualitative synthesis often requires iteration between synthesis and selection of primary studies. The reviewers need to report the general approach they planned to use for qualitative synthesis and the specific process they adopted during the conduct of the review. Just identifying a synthesis method by name is not sufficient for qualitative methods, for example, Eaves [47] identified (and synthesised) four different variants of *grounded theory*.

#### *Examples*

In their meta-analysis of the effectiveness of pair programming, Hannay et al. [21] provide a detailed discussion of the meta-analysis synthesis methods they adopted in Section 2.6. They introduced their discussion as follows:

> *We conducted separate meta-analyses for the three outcome constructs Quality, Duration, and Effort. Some studies applied several tests on the same outcome construct. In theses cases, we used the mean of the effect sizes over these tests to give only one effect size per outcome per study. Because we expected considerable heterogeneity, we decided to calculate the resulting meta-analytic effect sizes both under the assumption of the random-effects model and under the assumption of the fixed-effects model...*
>
> Quote 16: Discussion of Synthesis Method [21, p. 1112]

In their qualitative review of motivation, Beecham et al. [18] said:

> *We synthesised the data by identifying themes emanating from the findings reported in each accepted paper. These identified themes gave us the categories reported in our results section.*
>
> Quote 17: Discussion of Synthesis Method in Qualitative Review, [18, p. 863]

Continuing our running example of the comparison of single company and cross-company effort predictions:

> *Cross-company models do not have to outperform single company models to be beneficial. Therefore, investigated whether the estimate accuracy for single company projects obtained from cross-company models was* not significantly worse *than the estimate accuracy obtained from models developed from the single company project data.*
>
> *For this analysis, we used the* `R` *function* `binom.test` *to test whether data is consistent with the probability of Difference-2 being negative being 0.5. The* `binom.test` *provides confidence intervals on the observed probability of a negative Difference-2, as well as the probability that the true value of the probability of a negative Difference-2 is 0.5. We also investigated the sensitivity of the results by analysing what would happen if a new study found another incidence of a negative Difference-2 value.*
>
> Example 7: SynthesisMethod—Running Example Continued

### 3.17 Item 13e Methods Used for Sensitivity Analysis

PRISMA 2020 Definition: Authors should describe any sensitivity analyses conducted to assess robustness of the synthesised results

This item is not relevant for mapping studies.

Sensitivity analysis is about making comparisons between different ways of estimating the same effect. This contrasts with heterogeneity analysis that investigates whether different subgroups exhibit different effects. This item is standard for quantitative reviews and meta-analyses but, occasionally is also of relevance for qualitative reviews. For example, the authors might considers whether there are any identified themes or characteristics that would be removed if studies with high RoB were excluded.

#### *Explanation*

When sensitivity analysis is used authors should report sufficient details for readers to be able to assess the appropriateness of the analyses and to reproduce the reported results (with access to the data). Ideally, sensitivity analyses should be specified in the protocol, but unexpected issues may emerge during the review process that necessitate additional analyses.

Authors should report:

- Any sensitivity analyses that were performed, and details of each analysis (such as removal of studies at high risk of bias, use of an alternative meta-analysis model).
- Any sensitivity analyses that were not pre-specified should be identified.

*Examples*

Hannay et al [21] do not discuss the use of sensitivity analysis in the Methods section. They discuss their sensitivity analysis method when reporting their results as follows:

> *Fig. 2 shows one-study-removed analyses for each of the three outcome constructs. The plots show the meta-analytic effect size estimate when each study is removed from the meta-analysis. The resulting deviation from the full analysis indicates the sensitivity of the full analysis with respect to each study, that is, how much difference a given study makes to the meta-analysis.*
>
> Quote 18: Discussion of Sensitivity Analysis Method [21, Section 3.2, p. 1114]

Continuing our running example of the comparison of single company and cross-company effort predictions:

> *In order to avoid duplication, it was necessary to remove one of the two comparisons that used the cross-company Laturi database and the single company data set comprising 63 project (i.e., Study S1 and Study S6a). We chose to eliminate Study S1. We assessed the impact of this decision on our statistical analysis and the estimate of the median value of Difference-2.*
>
> *We also assessed the sensitivity of our statistical tests to the specific number of observed negative Difference-2 values by investigating the implication of a new primary study obtaining a negative Difference-2 value.*
>
> Example 8: Sensitivity Analysis Method—Running Example Continued

### 3.18 Item 13f Methods Used for Exploring Heterogeneity

PRISMA 2020 Definition: Authors should describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).

This item is not relevant for mapping studies. Although this item appears to apply only to quantitative reviews and meta-analysis, qualitative synthesis often aims to identify and investigate the implications of dis-confirming examples.

*Explanation*

For quantitative reviews and meta-analysis, if authors used statistical methods to explain differences in outcomes such as subgroup analysis or meta-regression, they need to provide enough details for readers to be able to assess the appropriateness of the selected methods and to reproduce the reported results (with access to the data).

Authors should report:

- The methods used.
- For subgroup analysis or meta-regression, or mixed effects analysis:
  - which factors were explored, levels of those factors, and which direction of effect modification was expected and why (where possible).
  - whether analyses were conducted using study-level variables (where each study is included in one subgroup only), within study contrasts (where data on subsets of participants within a study are available, allowing the study to be included in more than one subgroup), or some combination of the above.
  - how subgroup effects were compared.
- Other methods used and why they were adopted.
- Any analyses that were not pre-specified.

*Examples*

Ciolkowski [26] performed a subgroup analysis to support his meta-analysis inspection results. Although, it was reported in the discussion section it included the relevant information. The factors investigated were specified as follows:

> *Obvious candidates for such influence factors are: type of teams (nominal, real), experience of subjects (students, professionals), phase (requirements, design, code), and the control technique used (CBR or AR).*
>
> Quote 19: Data Used For Subgroup Analysis [26, p. 139]

Continuing our running example of the comparison of single company and cross-company effort predictions:

> *We thought it was likely that the relative accuracy of cross-company models and single company models might relate to the effort heterogeneity among the projects used to create the models.*
>
> *We assessed the relative heterogeneity of the single company data set compared with the cross company data set using the maxiumum and minimum effort values to calculate the effort range for each data set. If the information was reported, we used the maximum and minimum of the cross company projects used to construct the single company model (i.e., the cross company projects excluding the single company projects). Otherwise, we used the range values for the full cross company data set. We used the following statistic as a simple measure of single company project effort heterogeneity:*
>
> $$EffHet = \frac{MaxEff_{SC} - MinEff_{SC}}{MaxEff_{CC} - MinEff_{CC}} \quad (4)$$
>
> *A large* EffHet *value would suggest* high *heterogeneity for the single company project effort data. The* EffHet *value was compared with the Difference-2 value. If large* EffHet *values lead to inaccurate estimates (such that cross-company estimates are as good or better than the single company estimates), they should be associated with small or negative Difference-2 values.*
>
> Example 9: Investigating Heterogeneity—Running Example Continued

### 3.19 Item 14 Reporting Bias Assessment

PRISMA 2020 Definition: Authors should describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).

This item is not relevant for mapping studies.

*Explanation*

The validity of a synthesis may be threatened when the available results differ systematically from the missing results. This is known as "bias due to missing results"

and arises from "reporting biases" such as selective non-publication and selective non-reporting of results, but missing studies can also refer to lack of any studies addressing important evidence (such as the lack of any large-scale studies).

In the context of meta-analysis, there are statistical methods, such as funnel plots which aim to identify potentially missing results by plotting the effect size for each primary study against the effect size. For qualitative reviews and quantitative reviews that did not use meta-analysis, the usual approach is to use "tools" such as checklists that prompt users to consider symptoms of missing results. Page et al. [48] undertook a systematic review of such tools. They found four different tools (including GRADE [32]). The criteria associated with a *high* risk of bias summarised across all four tools is shown in Table 5. The GRADE criteria are shown in italicised text. The Recommendations column specifies how many of the four tools included the specific criterion. GRADE criteria are always used by a minimum of two tools because one of the other tools (NMA-Quality) uses exactly the same criteria as GRADE but in the context of network meta-analysis. The first five criteria were proposed by the AHRQ RRB tool [49], and the final three (together with two of the GRADE criteria) were proposed by the SAQAT Tool [50].

With respect to defining their method of assessing risk of synthesis bias, authors need to:

- Specify the methods (tool, graphical, statistical, or other) used to assess the risk of synthesis bias.
- Explain the process used to reach a judgement of overall risk of bias.
- If any adaptations to an existing tool were made (such as omitting or modifying items), specify the adaptations.
- If a new tool was developed for use in the review, describe the content of the tool and make it publicly accessible.
- Report how many reviewers assessed risk of synthesis bias, whether multiple reviewers worked independently, and any processes used to resolve disagreements between assessors.
- Report any processes used to obtain or confirm relevant information from study investigators.
- If a machine intelligence-based tool was used to assess risk of synthesis bias, report how the tool was used, how the tool was trained, and details on the tool's performance and internal validation.

#### *Practical Problems*

In SE secondary studies, discussion of the risk of missing data and/or primary studies has usually been based either on assessments that the threat of missing studies is low due to the intensity of the search process, or on reporting validation exercises that confirm that all (or a large percentage) of known studies were found by the search process.

A particular problem for authors of SRs is that they, themselves, are supposed to assess risk of synthesis bias based on the quality of the search and selection process they used for their own SR. This is extremely problematic for secondary studies that have not used meta-analysis. To minimize the obvious conflict of interest, we suggest using a checklist for assessing whether the search was as comprehensive and reliable as possible. Issues that can be considered include:

1) Whether all the limitations placed the search process were justified with respect to the study aims and research questions?
2) The rigour and transparency of the search process, considering the search method(s) used and their suitability given the goals of the secondary study.
3) How the search process was validated.
4) The rigour and transparency of the selection process.
5) How the selection process was validated.

The critical problem is one of converting these basic concerns into a set of questions related to separate domains that can be answered as objectively as possible, and which can allow risk of bias due to missing values across all the domains to be assessed. See [51] for specific questions related to the rigour of the search and selection process.

To assess the risk of synthesis bias once the papers have been selected and the data from the studies have been synthesised, criteria reported in Table 5 should be considered for secondary studies that have not performed a formal meta-analysis, in particular:

1) Whether some (otherwise eligible) papers/studies could not be accessed.
2) Discrepancies between published findings and grey literature findings.
3) The impact of small studies on the synthesis, particular a predominance of small, early, positive studies.

#### *Examples*

The SR reporting a comparison between single company and cross-company models [52] did not discuss reporting bias. For the purpose of our running example, Kitchenham applied the questionnaire shown in Table 6 to assess the reliability of the search and selection processes. This is not the most appropriate process. The questionnaire should be completed by at least two researchers, and all disagreements discussed and resolved. In the following example, we simply report the questionnaire and other information used to assess reporting bias.

> *The questionnaire shown in Table 6 was used to assess the reliability of the search and selection processes. The relationship between effect size and single company size was used to investigate whether there was any indication of missing data or missing studies.*

Example 10: Risk of Reporting Bias—Running Example Continued

### 3.20 Item 15 Certainty Assessment

PRISMA 2020 Definition: Authors should describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.

This item is not relevant for mapping studies.

TABLE 5: Criteria used to assess High Risk of Selective Publication Bias based on Four Checklists

| Criteria | Recommendations |
|---|---|
| *Evidence of funnel plot asymmetry (based on visual inspection of funnel plot or statistical test for funnel plot asymmetry)* | 4 |
| Smaller studies tend to demonstrate more favourable results (based on visual assessment, without funnel plot) | 1 |
| Decision would differ for estimates from a fixed-effect versus a random-effects model because the findings from a fixed-effect model are closer to the null | 1 |
| Substantial heterogeneity in the meta-analysis cannot be explained by some contextual or methodological factor | 1 |
| At least one study is affected by non-publication or non-accessibility | 1 |
| *Presence of small (often "positive") studies with for-profit interest in the synthesis* | 2 |
| *Presence of early studies (ie, set of small, "positive" trials addressing a novel method) in the synthesis* | 2 |
| *Discrepancy in findings between published and unpublished trials* | 3 |
| *Search strategies were not comprehensive* | 3 |
| *Methods to identify all available evidence were not comprehensive* | 2 |
| Grey literature were not searched | 1 |
| Restrictions to study selection on the basis of language were applied | 1 |
| Industry influence may apply to studies included in the synthesis | 1 |

TABLE 6: Risk of Bias Due to Missing Data (see Running Example 10)

| Item | Question | Assessment |
|---|---|---|
| 1 | Were all limitations placed the search process justified with respect to the study aims and research questions? | Yes (Low Rob), No (Serious Rob) |
| 2 | Was the process used to construct, refine and validate search strings for digital libraries appropriate? | Base assessment on sub-questions |
| 2.1 | Was the process used to construct the search strings explained and justified? | Process based on PICO or fully explained (Low Rob), Process inadequately explained (Moderate RoB |
| 2.2 | Was information about known primary studies used to refine search strings? | Yes (Low RoB), No (Moderate RoB) |
| 2.3 | Were search results from different digital libraries compared with one another and any discrepancies investigated? | Yes (Low RoB), No (Moderate RoB) |
| 2.4 | Were the digital libraries included in the search justified and sufficient to meet the study requirements and objectives in the light of any other search processes that were used ? | Yes (Low RoB), Probably Yes (Moderate RoB), Probably no (Serious RoB), No (Critical RoB) |
| 2.5 | Were all known papers found by the search strings? | Yes (Low RoB), No (Serious RoB) |
| 3 | Was the primary study selection process appropriate? | Base the assessment on the sub-questions |
| 3.1 | Did the selection process minimise researcher bias? | Two or more researchers assessed each citation (Low RoB), One assessed and another checked (Moderate RoB), One assessed and another checked a subset (serious Rob), One assessor (Critical RoB) |
| 3.2 | If the selection process was done in stages, were there clear criteria for stage completion | Yes (Low RoB), No (Moderate RoB) |
| 4 | How many (otherwise eligible) papers/studies could not be accessed | 0 (Low RoB), 1 (Moderate RoB), 2 (Serious RoB), 3 or more (Critical RoB) |

### *Explanation*

PRISMA 2020 expects authors to decide how certain (or confident) they are in the body of evidence for each important outcome. Such assessments are based on the properties of the set of primary studies that contribute to each outcome (i.e., a synthesis of primary study outcome data that addresses a specific research question) or finding (i.e., an observation arising from a qualitative assessment of primary study outcomes). This attempts to pull together information related to the methodological RoB of relevant primary studies (see Section 3.10) and risks associated with possible missing data (see Section 3.19) together with other issues such as the (in)consistency of findings across the set of relevant studies, (im)precision of study outcomes, and (in)directness of the study results.

The aim of the assessment is to assess the confidence in the evidence (also referred to as the *quality of the evidence*) supporting each finding on a four-point scale (High, Moderate, Low, Very Low).

For quantitative systematic reviews and meta-analysis, the GRADE tool attempts to perform such assessments [53] and is mentioned in the explanation and examples that Page et al. [1] provide to support PRISMA 2020. Furthermore, since some of the GRADE assessment criteria are used as part of the assessment of the risk of synthesis bias, it would seem sensible (but not mandatory) for authors who want to assess certainty to use the GRADE method.

For qualitative reviews, the GRADE-CERQual tool is available [54]. Currently, GRADE-CERQual considers four factors: methodological limitations, coherence, adequacy of data, and relevance. In [54] it is reported that inclusion of publication bias is "being explored".

Authors need to report:

- The method (i.e., tool, checklist, or system and version) used to assess certainty in the body of evidence.
- The factors considered (such as precision of the effect estimate, consistency of findings across studies) and the criteria used to assess each factor.
- The decision rules used to arrive at an overall judgement of the level of certainty (such as high, moderate, low, very low), together with the intended interpretation (or definition) of each level of certainty.
- Any review-specific considerations and any threshold used to assess imprecision and ranges of magnitude of effect that might be considered trivial, moderate or large, and the rationale for these thresholds and ranges.
- Any adaptions to an existing tool, together with the details and their justification.

- Any processes used to obtain or confirm relevant information from investigators.
- If a machine intelligence-based tool was used to support the assessment of certainty, how the automation tool was used, how the tool was trained, and details on the tool's performance and internal validation.
- The methods used for reporting the results of assessments of certainty, such as the use of Summary of Findings tables [53].

### *Basic GRADE and GRADE-CERQal Concepts*

Dybå and Dingsøyr [23] were the first SE researchers to discuss the GRADE approach to assess strength of evidence. They reported a quality checklist developed for an SR of agile methods. They made an important distinction between the quality evaluation of a study and an assessment of the overall strength of evidence associated with a topic of interest, when the topic may have been investigated using a variety of different empirical methods. However, the concept of assessing the strength of evidence has not been widely adopted by SE researchers. Budgen et al. [55] examined 49 reviews and found that only two made use of the GRADE approach to assess the strength of evidence of their findings (see Ali et al. [28] and Selleri Silva et al. [56]). Therefore, we provide a more detailed introduction to reporting strength of evidence based on the updated GRADE guidelines and the new GRADE-CERQual guidelines.

### *GRADE*

GRADE ( [32], [53]) was developed for two purposes:

1) rating quality of evidence in quantitative systematic reviews and guidelines,
2) grading strength of recommendations in guidelines.

In the context of systematic reviews, authors only need to rate the quality of the evidence.

A GRADE evaluation aims to evaluate each quantitative outcome from a systematic review on the following assessment scale [32]:

1) High quality - Further research is very unlikely to change our confidence in the estimate of effect.
2) Moderate quality — Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
3) Low quality - Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
4) Very low quality - Any estimate of effect is very uncertain.

A GRADE evaluation is a hierarchical evaluation process. At the top level, GRADE identifies five major criteria, all of which are assessed on a four-point scale (high, moderate, low and very low):

1) **Study Risk of Bias**: This should be based on the results of the assessment of methodological limitations (see Section 3.10) for the contribution made by the primary studies to each SR finding. Such assessments are usually based on assessing the risk associated with a number of different criteria, as shown in Table 7. Individual criteria contributing to study risk of bias are assessed on a subjective scale (usually using the same scale High, Moderate, Low and Very Low), assessing the likelihood that a specific criterion could have introduced bias. The highest risk level assigned to any of the individual criteria is used as the overall assessment criteria of study risk of bias.
2) **Risk of Synthesis Bias**: This is based on assessing the risk of synthesis bias (Section 3.19) due to publication bias. Such assessments are usually based on assessing the risk associated with a number of different criteria, as shown in Table 5. When multiple criteria are used, the subjective scale and method for obtaining an overall assessment are the same as those used for study risk of bias.
3) **Inconsistency**: This is the extent to which results are consistent for different studies. Indicators of low consistency are wide variation in mean effect sizes across different studies, non-overlapping confidence intervals for mean differences, low p-values for statistical tests of heterogeneity, or the proportion of the variance due to among study variation is large.
4) **Imprecision**: This concerns the length of confidence intervals for individual studies with quantitative outcomes. If the overall confidence interval from a meta-analysis of a set of primary studies included no effect in most cases, there would be no recommendation to adopt a new method. However, if overlap with the no effect condition is small and the costs of adopting the method are small, while the potential benefits are important, it may be appropriate to recommend the method.
5) **Indirectness**: This concerns the extent to which the studies represent the concerns of the practitioners or researchers who might be expected to make use of the results.

Imprecision and inconsistency are difficult to assess except in the context of formal meta-analysis, but indirectness can be assessed for any set of quantitative studies. For example, all of these issues can be used to assess indirectness for SE studies:

1) Lack of industry-based evaluations.
2) Lack of practitioner participants.
3) Studies dominated by a specific researcher or research group, with a lack of independent evaluations.
4) Limitations with respect to task difficulty (e.g., simple tasks capable of solution in a short time-period) and SE materials (e.g., unrealistically simple SE materials).
5) Excessive re-use of the same SE materials.

Many of these issues were identified as limitations applying to a set of studies in [26] and [25].

The overall GRADE assessment for a finding is based on the worst assessment of any of the contributing factors. However, the scales used for the contributing factors relates to *uncertainty* and so assess *lack of quality*. Thus, if any of the individual GRADE criteria are assessed as High, the corresponding assessment of certainty in (quality of) the body of evidence is Very Low.

### *GRADE–CERQual*

A GRADE–CERQual assessment specifies the level of confidence in a review finding using the terms:

1) High confidence: It is highly likely that the review finding is a reasonable representation of the phenomenon of interest.
2) Moderate confidence: It is likely that the review finding is a reasonable representation of the phenomenon of interest.
3) Low confidence: It is possible that the review finding is a reasonable representation of the phenomenon of interest.
4) Very low confidence: It is not clear whether the review finding is a reasonable representation of the phenomenon of interest.

GRADE–CERQual identifies four criteria that contribute to confidence in a qualitative review finding:

1) **Methodological Limitations**: This is discussed in [57]. It considers the extent to which there are concerns about the design or conduct of the primary studies. It can be based on the results of the assessment of methodological limitations (see Section 3.10) of the primary studies contributing to each SR finding.
2) **Coherence**: This is discussed in [58]. It considers how well-supported or compelling the fit is between the data from relevant primary studies and a specific finding.
3) **Adequacy of data**: This is discussed in [59]. It involves an overall determination of the degree of richness and quality of the data supporting a review finding.
4) **Relevance**: This is discussed in [60]. It considers the extent to which the body of evidence from the primary studies is applicable for the readers who might be expected to make use of it.

Each criterion is assessed using the terms:

1) No concerns, or very minor concerns regarding the specific criterion that are unlikely to reduce confidence in the review finding.
2) Minor concerns regarding the specific criterion that may reduce confidence in the review finding.
3) Moderate concerns regarding the specific criterion that will probably reduce confidence in the review finding.
4) Serious concerns regarding the specific criterion that are very likely to reduce confidence in the review finding.

Coherence can be assessed by considering three aspects:

1) Whether some data from a review contradicts a finding and there is no explanation of the contradiction.
2) Whether it is not clear if some of the data supports the finding, for example, if some of the underlying data is not well described or defined, or it seems likely that different primary studies have defined concepts in slightly different ways.
3) Whether there are plausible alternative descriptions, interpretations or explanations that could be used to synthesise the underlying data.

Adequacy of data is related to the number of studies from which each finding originated and the number and type of participants included in each relevant study.

Relevance in SE studies relates to :

1) The appropriateness and implications of the primary study eligibility criteria.
2) Information about the setting and context of each primary study, for example, the organisation size, the industry sector, and the type of software produced, see [61].

A GRADE–CERQual assessment for a specific finding usually equates to the lowest confidence level among the four criteria. Thus, if there are serious concerns about a specific criterion but no or low concerns about other criteria, the overall GRADE–CERQual assessment would be Very Low confidence in the finding.

GRADE–CERQual uses slightly different terminology to GRADE and includes three positive criteria rather than GRADE which considers only negative criteria.

### *Reporting GRADE and GRADE-CERQual Assessments*

GRADE and GRADE-CERQual both provide suggestions for reporting their assessments. Both discuss Evidence Profiles (EPs) which include a summary of the issues of concern for each major contributing factor, as well as the overall assessment (see [53] and [62]). GRADE also refers to Summary of Findings (SoF) Table, which only identifies the individual findings, and the overall assessment [62].

### *Examples*

There are no examples of SE systematic reviews using the most recent versions of GRADE or GRADE-CERQual. However, Dybå and Dingsøyr [23] discuss their use of an earlier version. For the purposes of planning their GRADE evaluation, they report that they mapped their eleven quality assessment questions (see Section 3.10) to four criteria:

1) Reporting: Questions 1 to 3.
2) Rigor: Questions 4-8.
3) Credibility: Questions 9 and 10.
4) Relevance: Question 11.

Thus, their assessment of the certainty of evidence was based solely on the study quality assessment criteria. In addition, their evaluation was based on all their primary studies not the individual SR findings.

The SR reporting a comparison between single company and cross-company models [52] did not discuss certainty assessment. For the purpose of our running example, Kitchenham performed a certainty assessment using the GRADE and GRADE-CERQual methods. This is not the most appropriate process. The assessment should have been undertaken by at least two researchers and all disagreements discussed and resolved. In the following example, we simply report the assessment criteria used in our running example:

> *The assessment of certainty in the body of evidence was made using the GRADE assessment criteria as follows:*
>
> - ***Study Risk of Bias**: This is based on the results of the assessment of methodological limitations (see the Examples subsection of Section 3.10) of the primary studies contribution to each SR finding.*
> - ***Risk of Synthesis Bias**: This is based on the results of the assessing the risk of synthesis bias, see the Examples subsection of Section 3.19.*
> - ***Inconsistency**: We investigated whether the results were consistent by identifying how many new studies favouring the cross-company model or single company model would be required to change the significance of the statistical analysis.*

- ***Imprecision**: The effect sizes used for summarizing the accuracy of effort estimation models do not have any accompanied estimate of variance. However, the we were able to calculate the confidence interval for the estimated probability that a cross-company model estimate would be more accurate than the single-company model estimate. The length of the confidence interval provides a means of assessing the (im)precision of the estimate.*
- ***Indirectness**: We discuss whether the primary studies considered the practical issues of using the results to assist effort estimation.*

*For qualitative findings, the GRADE-CERQual was used. For Methodological Weakness, the results of study risk of bias were considered in the context of the specific finding. For Coherence, the extent to which data from the individual studies were consistent with, or contradicted, the finding being assessed. For Data Adequacy, two issues were considered: whether there were possible missing primary studies and whether there were major gaps in the data obtained from the primary studies. For Relevance, the assessment was based on whether the finding identified issues of importance for software companies developing estimation models.*

Example 11: Certainty in the Body of Evidence—Running Example Continued

## 4 Results

If researchers have produced and trialled a comprehensive protocol, reporting the SR results should be fairly straightforward. The main practical problem is appropriately distinguishing Results from Discussion.

For mapping studies, the research questions should fully define the scope of the results and the reported results should directly address those questions.

For quantitative studies and meta-analyses, the Methods section should have identified all the analyses that will be reported. If the authors have performed additional unplanned analysis as a result of reading their primary studies and identifying unanticipated trends, such results can and should be reported, but they should be identified as resulting from unplanned analyses.

For qualitative reviews, there is often iteration between study selection and data analysis, so the protocol provides less of a constraint on synthesis than it does for mapping studies and quantitative reviews. Nonetheless, qualitative reviews will benefit from pre-specifying and trialling as much of their synthesis as possible, while being prepared to amend the selection process and the synthesis process if necessary. Any major changes to methods described in the protocol caused by the iterative nature of qualitative synthesis should be reported.

### 4.1 Item 16a Study Selection

PRISMA 2020 Definition: Authors should 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 .

#### *Explanation*

Authors need to report the outcomes of the selection process so that readers can understand the flow of retrieved records through to inclusion in the review. Such information is useful for future systematic review teams seeking to estimate resource requirements and for information specialists in evaluating their searches. Specifying the number of records found on a 'per database' basis will make it easier for others to assess whether they have successfully replicated a search.

Specifically authors should report:

- Ideally using a flow diagram, the number of:
  - records identified;
  - records excluded before screening (for example, because they were duplicates or deemed ineligible by machine classifiers);
  - records screened;
  - records excluded after screening titles or titles and abstracts;
  - reports retrieved for detailed evaluation;
  - potentially eligible reports that were not retrievable;
  - retrieved reports that did not meet inclusion criteria and the primary reasons for exclusion (such as ineligible study design, ineligible population);
  - number of studies and reports included in the review.
- For updates, the number of studies and reports included in the previous review.
- If automated tools were used, how many studies were excluded by tools and how many by humans.

#### *Examples*

Kitchenham and Brereton [9] provided a detailed narrative description of their selection process, which would have been much easier to understand had it been accompanied by the flow diagram, shown in Figure 1. In their Appendix C, they list all included papers indicating which were duplicate reports and identifying papers that reported multiple studies.

### 4.2 Item 16b Identify Near-Miss Studies

PRISMA 2020 Definition: Authors should cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.

#### *Explanation*

Authors should identify any studies that might appear to meet the inclusion criteria, but which were excluded, and explain the reason for the exclusion. Recording this information allows readers to make an assessment of the validity and applicability of the systematic review.

#### *Examples*

In Appendix B. of their systematic review of systematic review process research in SE, Kitchenham and Brereton [9] reported the details of 10 papers that were eliminated during data collection, together with the reason for the exclusion.

### 4.3 Item 17 Study Characteristics

PRISMA 2020 Definition: Authors should cite each included study and present its characteristics

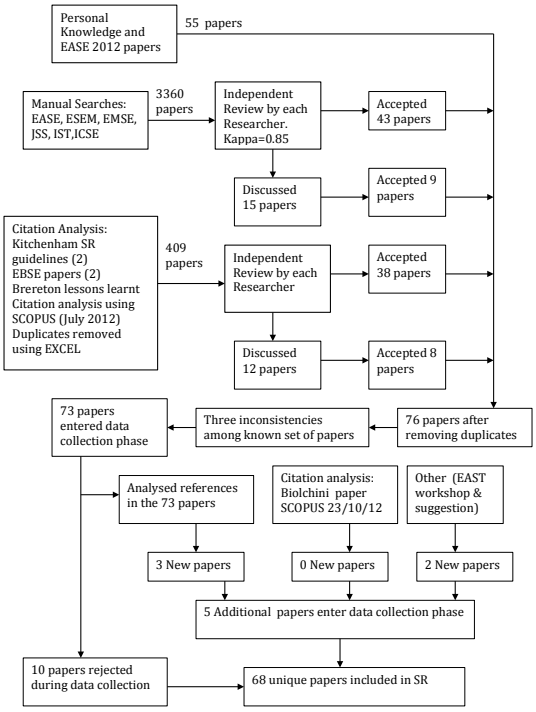


Fig. 1: Flow diagram of the Selection Process Reported in [9]

### *Explanation*

Readers need to know the characteristics of the studies that have contributed to the review. Characteristics of interest might include study design features, characteristics of participants, how outcomes were measured. Tabulating or graphing such characteristics allows comparisons between primary studies. Citing each primary study allows readers to retrieve reports they consider important.

### *Examples*

Tabulating information about the primary studies included in a secondary study is required for all types of secondary study.

Kitchenham et al. reported basic information about each systematic review they included in a systematic review of SRs (which would now be referred to as a tertiary study) in Table 2 of their paper [12]. They provided citation information and information about the topic addressed by the SRs (referred to as systematic literature reviews (SLRs) in the article), the systematic review references cited, whether the SR included practitioner-relevant guidelines and the number of primary studies included in the review. Some of the studies classified as SRs would now be classified as mapping studies, and study S8 would now be considered an update of S7.

## 4.4 Item 18 Risk of Bias in Studies

PRISMA 2020 Definition: Authors should present assessments of risk of bias for each included study.

This is not usually required for mapping studies.

### *Explanation*

For readers to understand the internal validity of a systematic review's results, they need to know the risk of bias in the results of each included study. Reporting summary data alone, such as a quality score, is inadequate because it fails to inform readers about which studies had each particular methodological shortcoming.

Authors should:

- Present tables or figures indicating the risk of bias in each domain/component/item assessed and an overall study-level risk of bias.
- Provide a justification for each risk of bias judgement, for example, relevant quotations from the primary study reports.

### *Examples*

Continuing the running example of the comparison between single company and cross-company models:

> *The basic risk of bias for each study in comparison of single company and cross-company estimation models is shown in Table 7. The RoB starting point for each study is* Low *because none of the studies are randomised trials. However, because the industry datasets are the data sources that would be used by single companies to assist their effort estimation, we have revised our initial assessment to* Moderate. *Although Study 6 analysed six different single company data sets, it used exactly the same basic method for each comparison, so we provide only one risk of methodological RoB assessment for that primary study.*
>
> *We excluded Study 1 and Study 7 from our synthesis because both studies were duplicated by other studies (Study 6 and Study 2 respectively) that used the same cross-company and single company data sets and, in addition, Study 2 and Study 6 had fewer methodological problems. Study 1 was unable to calculate estimates for some of the single company data sets using the single company data. Study 7 used a different modelling approach and only based predictions on the most recent 15 data points in the single company data set. Neither study reported median magnitude relative error which was reported by all other studies.*

Example 12: Risk of Bias Assessment in Studies—Running Example Continued

## 4.5 Item 19 Results of Individual Studies

PRISMA 2020 Definition: For all outcomes, authors should 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

This item applies only to quantitative systematic reviews and meta-analyses.

### *Explanation*

Presenting data from individual studies both clarifies each study's contribution to the findings and supports reuse of the data by others seeking to perform additional analyses or perform an update of the review. There are different ways of presenting results of individual studies (such as tables, or Forest plots). Visual display of results supports interpretation by readers, while tabulation of the results makes it easier for others to reuse the data. Ideally, authors should report primary study data both visually and in tables.

For each primary study, authors should report:

- all outcomes summary statistics for each group (for example, the control group and treatment group),

TABLE 7: Risk of Bias Due to Methodological Limitations in Studies Comparing Single Project and Cross Project Effort Estimates (see Running Example 12)

| Criteria | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Was the data analysis process appropriate?** | | | | | | | | | | |
| Was data investigated for outliers and distribution? | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes |
| If Yes, were the results used appropriately? | Yes | Yes | Yes | Yes | Yes | Yes | NR | Yes | Yes | Yes |
| Were statistical tests appropriate? | No test | Yes | Yes | Yes | Yes | Yes | No test | Yes | Yes | Yes |
| **Sensitivity or residual analysis?** | | | | | | | | | | |
| Did the resulting estimation models subject to sensitivity or residual analysis? | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes |
| Was the result of the sensitivity or residual analysis used to refine models or to explain results? | Yes | Yes | Yes | Yes | Yes | Yes | NR | Yes | Yes | Yes |
| **Were accuracy statistics based on the raw data?** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

- any reported effect sizes and their confidence intervals.

*Examples*

In their meta-analysis of pair programming, Hannay et al. [21] provide a display of the effect sizes and effect size confidence intervals for each study organized according to the output metric (i.e., Quality, Duration, and Effort) which includes both the numerical values and a Forest plot.

Continuing our running example of the comparison between single company and cross-company models:

> *The results found by each primary study are summarised in Table 8, where* CCEffect *is the MdMRE for single company predictions based on the best fitting cost estimation model,* CCmethod *specifies the method used to obtain the best fitting cross-company model, and* SCEffect *is the MdMRE value using the same model building method on the single company data,* Difference-1 *is the difference between* CCmethod *and* SCEffect. BestSCEffect *is the MDMRE value obtained from the best fitting model derived from the single company data.* Difference-2 *is the difference between* CCmethod *and* BestSCEffect. EffHet *is the project effort range of the single company project data in comparison with the project data effort range of the cross-company project data. The studies are presented in publication date order.*

Example 13: Individual Study Results—Running Example Continued

### 4.6 Item 20 Results of Analyses and Syntheses - General Issues

Results of analysis and synthesis will be very different for different types of secondary studies.

For quantitative studies (particularly meta-analysis), authors should report their results using the graphical and synthesis methods defined in the Methods section. They need to consider the reporting requirements for each sub-item.

For mapping studies, authors should report the characteristics of each primary study and the results of the analysis they performed to address each research questions. Sometimes they can report the basic primary study characteristics in a single table; sometimes, however, characteristics differ between different types of primary study. Items 20a and 20b are relevant for most mapping studies, the other sub-items are not.

Qualitative studies will need to report the details of how they performed their syntheses, as well as the results of the synthesis. In particular, if authors have developed novel models that are more complex than simple identification and definitions of themes and relationships, the details of the synthesis activity and any validation of the resulting models will need to be reported in accordance with process specified in the Methods Section.

### 4.7 Item 20a Characteristics and Risk of Bias

PRISMA 2020 Definition: For each synthesis, authors should briefly summarise the characteristics and risk of bias among contributing studies.

For mapping studies, usually only study characteristics are reported.

*Explanation*

Authors of all types of secondary study need to summarise the characteristics of the included primary studies, although for secondary studies that have a variety of different research questions (particularly mapping studies), this may be reported on a per-research question basis. For quantitative SRs and qualitative reviews, this may be organised on a per-outcome basis.

For secondary studies other than mapping studies, some assessment of risk of bias should have been performed and this should also be reported. Providing a brief summary of the characteristics and risk of bias among studies contributing to each synthesis (meta-analysis or other) should help readers understand the applicability and risk of bias in the synthesised result.

*Examples*

Continuing the running example of the comparison between single company and cross-company models:

> *The general characteristics of the cross-company and single company models are shown in Table 9. It is clear that Study S2 has used the same cross-company dataset and single company dataset as Study S6a. However, the selection criterion for the cross-company data used by S2 was unclear while Study S6 always used all the cross-company data except that of the single company being estimated. For this reason, we kept study S6a in the data synthesis and omitted Study 2. The selection for criteria for determining the single company dataset seems well-founded and unbiased for all primary studies. However, only Study S4 used a single data set completely independent of the cross-company data set.*

Example 14: Characteristics and Risk of Bias—Running Example Continued

TABLE 8: Individual Study Results (see Running Example 13)

| Study | CC Effect | CC Method | SC Effect | Difference-1 | Significant | Best SC Effect Size | Best SC Method | Difference-2 | EffHet |
|---|---|---|---|---|---|---|---|---|---|
| S2 | 46.0 | CART | 46.2 | -0.2 | No | 41 | SWR | 5 | 0.997 |
| S3 | 32.0 | OLS | 34.0 | -2.0 | No | 26.0 | ANOVA-e | 6.0 | 0.007 |
| S4 | 38.0 | OLS | 27.0 | 11.0 | Yes | 27.0 | OLS | 11.0 | 0.229 |
| S5 | 68.3 | ROR | 26.3 | 42.0 | Yes | 17.8 | CART_p | 50.5 | 0.018 |
| S6a | 46.0 | Analogy | 39.0 | 7.0 | No | 39.0 | Analogy | 7.0 | 0.995 |
| S6b | 13.0 | ANOVA | 23.0 | -10.0 | Yes | 20.0 | Analogy | -7.0 | 0.087 |
| S6c | 32.0 | Analogy | 37.0 | -5.0 | No | 22.0 | Analogy | 10.0 | 0.378 |
| S6d | 30.0 | OLS | 55.0 | -25.0 | Yes | 25.0 | Analogy | 5.0 * | 0.791 |
| S6e | 31.0 | Analogy | 32.0 | -1.0 | No | 32.0 | Analogy | -1.0 | 0.270 |
| S6f | 30.0 | ANOVA | 26.0 | 4.0 | No | 26.0 | OLS | 4.0 | 0.403 |
| S8 | 44.4 | MSWR | 23.4 | 21.0 | Yes | 23.4 | MSWR | 21.0 | 0.353 |
| S9 | 62.0 | SWR | 38.0 | 24.0 | Yes | 38.0 | SWR | 24.0 | 0.035 |
| S10 | 61.0 | SWR | 60.0 | 1.0 | No | 60.0 | SWR | 1.0 | 1.0 |

TABLE 9: Characteristics and Risk of Bias of Studies Comparing Single Company and Cross-Company Effort Estimation Models (see Running Example 14)

| Study | Cross Co Dataset | Cross Co Total Projects | Cross Co Model Num Projects | Single Co Data set | Single Co Num Projects | Single Co Selection Criteria | Methodological ROB |
|---|---|---|---|---|---|---|---|
| S2 | Laturi | 206 | 143 | Laturi | 63 | Largest single company | Moderate |
| S3 | ESA | 166 | 131 | ESA | 29 | Largest single company | Moderate |
| S4 | ISBSG | 451 | 145 | Megatec | 19 | Single company data | Moderate |
| S5 | ISBSG | 324 | 310 | ISBSG | 14 | Largest single company | Moderate |
| S6a | Laturi | 206 | 143 | Laturi | 63 | Single company10+ | Moderate |
| S6b | Laturi | 206 | 193 | Laturi | 13 | Single company10+ | Moderate |
| S6c | Laturi | 206 | 194 | Laturi | 12 | Single company10+ | Moderate |
| S6d | Laturi | 206 | 195 | Laturi | 11 | Single company10+ | Moderate |
| S6e | Laturi | 206 | 196 | Laturi | 10 | Single company10+ | Moderate |
| S6f | Laturi | 206 | 196 | Laturi | 10 | Single company10+ | Moderate |
| S8 | Tukutuku | 53 | 40 | Tukutuku | 13 | Largest single company | Moderate |
| S9 | Tukutuku | 67 | 53 | Tukutuku | 14 | Different single company | Moderate |
| S10 | ISBG | 872 | 680 | ISBG | 187 | Largest single company | Moderate |

### 4.8 Item 20b Analysis or Synthesis Results

PRISMA 2020 Definition: Authors should present results of all statistical syntheses conducted. If meta-analysis was performed, 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.

Qualitative studies will report the results of qualitative synthesis.

#### *Explanation*

Authors of any secondary study need to report the results of all forms of analysis and synthesis identified in the methods section. For mapping studies, this should be the analysis (including graphs and tables) used to answer each research question. For qualitative reviews, this will involve answers to specific research questions and/or the development of any new conceptual model. For quantitative studies that *do not* involve meta-analysis this may result in a qualitative-style analysis based on discussing any trends in the outcomes from the primary studies visible in graphs or tables comparing favourable and unfavourable outcomes against study characteristics. For quantitative studies using meta-analysis, authors should present the meta-analysis results obtained for each effect size. If authors of meta-analyses undertook additional analysis/synthesis in addition to their planned analyses/syntheses, these should be reported, but the authors need to make it clear that the analyses/syntheses were not pre-planned.

#### *Examples*

In their meta-analysis, Hannay et al. [21] provide the overall fixed and random model meta-analysis results in their Forest plot (see their Fig.2).

In their qualitative review of software engineer motivation, Beecham et al. [18] report the answers to each research question separately, but also considered the relationship between their research questions (see their Fig. 6). They also considered the structure of the responses. For example for RQ1 *What are the characteristics of software engineers*, they identified 43 papers that addressed the issue, from which they identified 24 attributes which they tabulated (together with the papers that reported them) in their Table 5. However, they also say:

> *However, a closer inspection shows that these attributes can be structured into three linked categories. The first category contains the 'raw' characteristics of Software Engineers. The second contains factors that control whether or not a particular individual will have those characteristics. The third contains moderators which determine the strength of a characteristic within an individual.*

Quote 20: Characteristics of Software Engineers [18, p. 866]

They point out that, as suggested by the wider literature, individual software engineers have individual characteristics profiles that can change over time.

They also discuss the impact of Control factors that relate to an individual personality and strengths and weaknesses, and the impact of career stage and culture. They note that these issues are often mentioned as moderators in the literature and are likely to moderate the strength of each characteristic a software engineer possesses. They conclude that:

> *Our findings suggest that an engineer's personality, career path preference and competencies will control whether each of the 16 characteristics listed in Table 5 form part of his or her make-up.*
>
> Quote 21: Impact of Control Factors [18, p. 867]

Two of the goals of realist qualitative synthesis are theory building and testing [5]. Both of these issues need to be reported in the results section, and in the case of software engineering motivation they were addressed in a subsequent paper by the same group of researchers that used the results of their qualitative review to build a new integrated model of software engineering motivation [19] and evaluate it against other models of motivation. They started from the relationship between their research questions and evolved the model according to the detail of the SR findings. They describe the first part of their model building process as follows:

> *The software engineer characteristics listed in Table 2 fall into two different categories: characteristics of the individual, and expressed needs. So for example, the literature says that a software engineer is introverted by nature, but also has a need for variety in his/her work. Re-presenting this set of results by clustering characteristics on one side and needs on the other, gives a picture as shown in Fig. 2.*
>
> Quote 22: A New Model of Software Engineering Motivation [19, p. 222]

They discussed each of their four research questions and explain how their findings with respect to each research question allowed them to build their MOCC (Motivators, Outcomes, Characteristics and Context) model.

Having derived their new model, Sharp et al. [19] then compared it with other motivation models in the literature saying:

> *In doing so, we specifically focus on how the models in the literature relate to the components of the new derived model. For example, does an existing model provide more detail about a particular component, or does it explore the relationship between elements of different components, or does it offer a perspective across all components (or indeed identify other components). We want to highlight what is missing from the new model so that it can be enhanced.*
>
> Quote 23: Model Evaluation Criteria [19, p. 223]

Sharp et al. [19] summarise the scope and results of their model evaluation activity in terms of potential missing factors and potential missing relationships. After discussing the differences between the MOCC model and other models

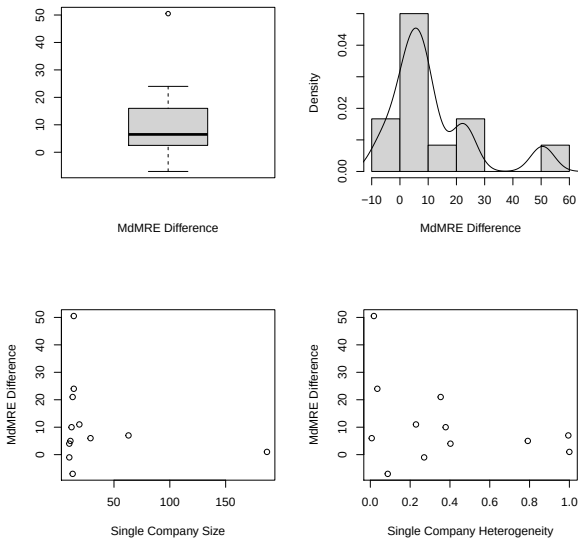


Fig. 2: The Distribution of the Difference in MdMMRE for the Best Fitting cross-company and Single Company Model Predictions (see Running Example 15)

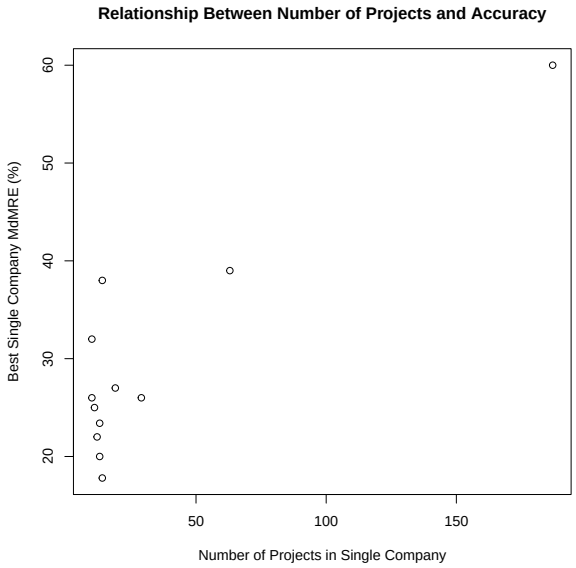


Fig. 3: The Relationship between the Number of Projects in the Best Fitting Single Company Model and Prediction Accuracy (see Running Example 15)

in more detail, Sharp et al. conclude that motivation is context dependent and the literature does not shed much light on how this can be represented.

Continuing the running example of the comparison between single company and cross-company models:

> *Table 8 reports the statistical significance of the* Difference-1 *in the column labelled* Significant. *This indicates that two studies (S6b and S6d) found the cross-company model significantly more accurate than the single company model while four studies (Study S3, Study S4, Study S8 and Study S9) found the single company model significantly better than the cross-company model. Seven*

*studies found no significant difference.*

*A problem with this assessment is that, as shown in the column labelled* BestSCEffect*, the best fitting cross-company model will not necessarily result in the best single company predictions. Comparing the best fitting single company model predictions with the best fitting cross-company model predictions only two results favoured the cross-company model. The largest negative difference was -7 which would probably not be significant (since a value 7 was not significant for Study S6a).*

*The upper two panes of Figure 2 show the box plot and kernel density plots of the* Difference-2 *values for predictions based on the best cross-company and single company models. The median value of the MdMMRE difference values is 6.5% with only two of the 14 differences being less than zero. Thus, the frequency of results favouring the cross-company model is 2/12=0.17 with 95% confidence interval bounds (0.02, 0.48). If the true probability was 0.5 (i.e., there was no difference between the accuracy of the best cross-company model and the best single company model) the probability of obtaining only two negative values is* $p = 0.039$*. Thus, we can reject the hypothesis that the best estimates from the cross-company model are no worse than the best estimates from the single company model (*$p < 0.05$*).*

*The lower two panes investigate the relationship between MdMMRE difference values and single company size and single company effort heterogeneity. Two of the single company data sets included more than 50 data points and they also exhibited high levels of effort heterogeneity with a range of project effort values very close to the range of project effort values in the cross-company data set.*

*We also considered the relationship between model accuracy and model size for the single company models. This analysis was not planned when the study protocol was developed, but observing the data shown in Table 8, it seemed that, in contrast to normal statistical assumptions, large single company data sets did not lead to more accurate single company effort estimation models. A scatter plot of the number of projects in each single company data set and the percentage MdMRE of the best fitting single company estimation model for the company is shown in Figure 3. Since, the larger the value of percent MdMRE the worse the accuracy of the single-company model, Figure 3 shows that 8 of the 10 single company models with less than 30 projects had MdMRE values less than 30% and two of the four models with MdMRE greater than 30% corresponded to the two largest single company data sets. The remaining two studies had 14 and 10 projects respectively with MdMRE values of 38% and 32%.*

Example 15: Synthesis Results—Running Example Continued

### 4.9 Item 20c Sensitivity Analysis

PRISMA 2020 Definition: Authors should present results of all sensitivity analyses conducted to assess the robustness of the synthesised results.

This item is irrelevant for mapping studies.

For other systematic reviews, it is mainly concerned with identifying whether any findings are dependent either on specific primary studies, or primary studies of a particular type. It is commonly used to support meta-analysis but it may also be used in the context of qualitative synthesis.

#### *Explanation*

Presenting results of sensitivity analyses conducted allows readers to assess how robust the synthesised results were to decisions made during the review process. Reporting results of all sensitivity analyses is important; presentation of a subset, based on the nature of the results, risks introducing bias due to selective reporting.

Authors should report all sensitivity analyses that were conducted and comment how robust the overall synthesis was given the results of all the sensitivity analyses.

In the context of meta-analysis, sensitivity analysis is usually based on reanalysing the data leaving one primary study out at a time. Meta-analysis tools usually provide automated methods for doing such analyses and producing forest plots of the results that include standard statistical summary information (i.e., estimates of the effect size, confidence intervals, P values and heterogeneity measures)[3].

#### *Examples*

Table 3 of the meta-analysis of pair programming by Hannay et al. [21] shows a leave-one-out Forest plot showing the revised effect size after omitting each project in turn, for each of the three outcome measures (Quality, Duration and Effort). They discuss the sensitivity analysis as part of their discussion of the overall meta-analysis. For example, their discussion of the effects for Quality is as follows:

*The three studies by Domino et al. (2007), Arisholm et al. (2007), and Madeyski (2006) contribute more than 50% of the total weight in the meta-analysis for Quality. The one-study-removed analysis shows that the meta-analysis is most sensitive to the inclusion/exclusion of Williams et al. (2000). Heterogeneity is significant at a medium level (*$Q = 35.97$*;* $p < 0.01$ *;* $I^2 = 63.86\%$*).*

Quote 24: The Impact of Pair Programming on Quality [21, p. 1114]

Continuing the running example of the comparison between single company and cross-company models:

*If we had selected Study S1 for inclusion rather than Study S6a, the median MdMRE difference would have been reduced from 6.5 to 5. Since the direction of the effect is not changed the results of the binomial test would have been also have been unchanged. Overall, choice of study to include had only a minor effect on the results but including Study S6a is consistent with the basic protocol of comparing the best cross-company model with the best single company model.*

*We tested the sensitivity of our statistical analysis to our specific set of primary studies by assessing the impact on our analysis if we found a new study exhibiting a negative*Difference-2*. In this case, we would have 13 projects with 3 negative* Difference-2 *values. However, after failing a test that the underlying probability of a negative* Difference-2 *was 0.5, subsequent tests with additional primary studies should be based on the one-sided test that the probability is less than 0.5. Using a one-sided test, the estimate of the probability of a negative* Difference-2 *value would be 0.231 with 95% confidence interval limits (0.000, 0.49). The probability that the true underlying probability of a negative* Difference-2 = 0.5 *would be 0.046. It would take two new primary studies both with negative* Difference-2 *values to produce data that would reject the hypothesis that the true probability of a negative* Difference-2 *was less than 0.5. This suggests that the finding that the single company estimates are significantly more accurate than the cross-company estimates is moderately robust.*

Example 16: Sensitivity Analysis Results—Running Example Continued

3. Analysis of the differences between studies of low and high risk of bias would be classified as heterogeneity analysis.

### 4.10 Item 20d Investigation of Heterogeneity

PRISMA 2020 Definition: Authors should present results of all investigations of possible causes of heterogeneity among study results.

For meta-analysis, exploratory formal statistical methods are available to investigate heterogeneity. For quantitative reviews that do not use formal meta-analysis, heterogeneity assessments are based on looking for study characteristics that seem to influence outcomes. For qualitative reviews, contrasting viewpoints and apparent contradictions should be identified and reported as part of the data synthesis and as part of any model-building exercise (see Quote 22).

#### *Explanation*

It is important to report all investigations of possible causes of heterogeneity among study results. Such studies:

1) help users to understand which factors *may, or may not,* explain variability which may, in turn, influence decision making.
2) help researchers to generate hypotheses that can be tested in future studies

Selective reporting of heterogeneity (e.g., only reporting effects that were statistically significant rather than all effects that were tested) may mislead users and researchers.

If informal methods (that is, those that do not involve a formal statistical test) were used to investigate heterogeneity (which may arise particularly when the data are not amenable to meta-analysis) authors should describe the results they observed. For example, present a table that groups study results by outcome group (i.e., those that favoured the control and those that favoured the alternative) and/or overall risk of bias and comment on any patterns observed.

Authors who perform formal heterogeneity analysis, should:

- present the results regardless of the statistical significance, magnitude, or direction of effect modification.
- identify the studies contributing to each subgroup identifying whether the subgroups were based on classifications applied at the primary study level, or by splitting data within studies into subgroups (for example, student or practitioner participants).
- report results with due consideration to the observational nature of the analysis and risk of confounding due to other factors.
- report the exact P levels for all tests (which will depend on the analysis methods used), the standard error and confidence/credible interval of effect sizes and measures of heterogeneity.

#### *Examples*

If informal methods of heterogeneity analysis are used, they are usually based on tabulating the relationship between contextual study factors and study outcomes and providing textual discussion of the tables and their implications (i.e., a form of narrative synthesis). For example, Jørgensen [27] investigated whether the accuracy of formal prediction models and prediction made by experts depended on whether the models were highly calibrated or whether the experts had more information than the models. He presented the assessment of evidence concerning model calibration in Table 2 of his paper. He reported that there appeared to be weak evidence that related expert performance to the level of model calibration. He also discussed the two studies that provided counter evidence as follows.

> *A discussion with the author of Study 14 suggests that a possible reason for the model's performing well in spite of the low calibration may have been that the set of projects that led to the construction of the estimation model was similar to the set of projects on which the model was applied.... The "mixed evidence" of the models with a low level of calibration in Study 2 is caused mainly by one expert who provided extremely inaccurate estimates ...*
>
> Quote 25: Model Calibration and Accuracy [27, p. 459]

Jørgensen [27] presented his assessment of the impact of experts having additional information in Table 3 of his paper and commented:

> *...the majority of the studies were based on providing different inputs to the experts than to the models, which is what actually happens in real life software development contexts. Only four studies provided the same information to the models and the experts. Hence, it is difficult to draw conclusions about the importance of contextual information for the relative estimation performance of experts and models based on Table 3 alone.*
>
> Quote 26: Additional Expert Information and Accuracy [27, p. 459]

In contrast, reporting the heterogeneity investigation applied to a meta-analysis is more straightforward. Ciolkowski [26] reported his heterogeneity results in a single table (Table 2). The columns identified each subgroup, the standardised effect size difference and its confidence interval, the P value, the $I^2$ heterogeneity value, the number of projects in each subgroup and the percentage of variance explained by the corresponding moderator variable. The rows specified the detailed statistics for each subgroup analysis and provided the overall results as a baseline to assess the impact of the subgroups. He also discussed the results in terms of possible confounding effects, for example, in his discussion of the largest observed effects, he said:

> *Surprisingly, we observed the largest effect sizes for the subgroups comprising professional developers and nominal teams. This could be taken to mean that professional developers perform better than students (i.e., PBR has a higher advantage for professional developers), and that nominal teams are not comparable with real teams, which questions the basic assumption that there is no difference between nominal and real teams in terms of their effectiveness.*
>
> Quote 27: The Results of Heterogeneity Analysis [26, pp. 139–140]

Ciolkowski suggests that a possible reason for this observation was confounding with other variables.

Continuing the running example (see Example 15) of the comparison between single company and cross-company models [7]:

*We expected that:*

1) *Single company data sets exhibiting a range of effort values similar to that of the cross-company data set, would produce produce estimation models that had similar accuracy to the cross-company models.*
2) *Single company data sets that had a very restricted range of effort values compared to the cross-company data set would produce estimations models that were much more accurate than any cross-company model.*

*Both these conjectures had some support:*

1) *Study S10 and Study S6a had large single company data sets (63 and 187, respectively) with ranges of effort values close to the range of values in the cross-company data set (effort heterogeneity values of 1 and 0.995 respectively) and had similar relatively poor levels of model accuracy for the cross-company and single company models (for S10 the MdMRE values were 61 and 60 for the cross-company and single company models respectively, for S6a the MdMRE values were 46 and 39 respectively).*
2) *Study S5 and Study S9 both had small ranges of effort values (i.e., effort heterogeneity of 0.018 and 0.035 respectively) and had single company models more accurate than cross-company models.*

*However Study 6b contradicted both conjectures. It was a small data set, and, also, had a low effort heterogeneity (0.087). However, $MdMRE = 13$ for the cross-company model. This was the most accurate of all the cross-company models, and substantially outperformed the single company model.*

Example 17: Factors Influencing Outcomes—Running Example Continued

### 4.11 item 21 Reporting Biases

PRISMA 2020 Definition: Authors should present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.

This item is irrelevant for mapping studies.

#### *Explanation*

Presenting assessments of the risk of bias due to missing results in syntheses allows readers to assess potential threats to the trustworthiness of a systematic review's results. Providing the evidence used to support judgements of risk of bias allows readers to determine the validity of the assessments.

Although there are statistical methods that can be used to assess the extent of publication bias for meta-analyses, other methods suitable for other types of SR are identified in Table 5.

Authors should:

- Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.
- If a tool was used to assess risk of bias due to missing results in a synthesis, present responses to questions in the tool, judgements about risk of bias, and any information used to support such judgements to help readers understand why particular judgements were made.

In the context of meta-analysis, funnel plots are often used to assess the risk of missing data. Authors reporting funnel plot results should:

- Present the plot and specify the effect estimate and measure of precision used in the plot (presented typically on the horizontal axis and vertical axis respectively).
- If funnel plot asymmetry was tested, report the exact P value and other relevant statistics.
- Report the results of any sensitivity analysis investigating the potential impact of missing values.
- If there might have been selective non-reporting, identify the studies with probable missing values and consider their impact.

#### Examples

Hannay et al. [21] report funnel plots for each of their three outcome measures (see Fig. 3, Fig 4 and Fig 5.). They also report the impact of *imputing* the missing project values[4].

Hannay et al. point out that the imputation method should only be used for sensitivity analysis since it is not possible to be sure that asymmetry is *caused* by missing values. For example, they report an imputed value that they suggest might not be due to a missing value but might be due to an overly large effect size on the opposite side of the funnel.

Continuing the running example of the comparison between single company and cross-company models:

*The risk of missing data questionnaire is shown in Table 10. The main weakness is that the second search based on removing date limitations was not organized and reported as rigorously as the first search which suggest that the assessment should be Moderate. However, the additional search processes included both backwards snowballing and approaches to leading researchers, and no additional candidate primary studies were found. Thus, we assess the likelihood of missing primary studies to be Low. We evaluated all other criteria as have Low or Very Low RoB, we conclude that the overall Risk of Bias due to Missing data should be considered Low.*

Example 18: Risk of Missing Data—Running Example Continued

### 4.12 Item 22 Certainty of Evidence

PRISMA 2020 Definition: Authors should present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.

This item is irrelevant for mapping studies.

#### *Explanation*

Authors should report their assessment of the certainty in (or quality of) the body of evidence for each outcome or finding. It is useful to report both the level of certainty in the evidence and the basis for the assessment. Summary tables such as the Evidence Profiles (EP) or Summary of Findings

4. Points on a funnel plot represent the effect size for each project (on the x-axis) and its precision measures as the inverse standard error on the y-axis. Results for a specific precision should be symmetric about the overall average effect size and studies with high precision should be close to the average while studies with low precision will exhibit greater variance. Missing values are imputed by adding points to increase the balance of the funnel plot, until tests of significance no longer detect any significant lack of balance. Then the overall mean is recalculated by including the imputed values in the meta-analysis.

TABLE 10: Assessment of Risk of Bias Due to Missing Data (see Running Example 18)

| Item | Question | Comment | Assessment |
|---|---|---|---|
| 1 | Were all limitations placed the search process justified with respect to the study aims and research questions? | Yes | Low RoB |
| 2 | Was the process used to construct, refine and validate search strings for digital libraries appropriate? | Base assessment on sub-questions | Low RoB |
| 2.1 | Was the process used to construct the search strings explained and justified? | Process based on PIO and fully explained | Low RoB |
| 2.2 | Was information about known primary studies used to refine search strings? | Yes | Low RoB |
| 2.3 | Were search results from different digital libraries compared with one another and any discrepancies investigated? | Reported in Conference paper [6] | Very low RoB |
| 2.4 | Were the digital libraries included in the search justified and sufficient to meet the study requirements and objectives in the light of any other search processes that were used ? | Six digital libraries (including general indexing libraries and computer science digital libraries), backwards snowballing, seven specialized conferences, authors contacted | Very low RoB |
| 2.5 | Were all known papers found by the search strings? | All except unpublished paper | Low RoB |
| 3 | Was the primary study selection process appropriate? | Base the assessment on the sub-questions | Low RoB |
| 3.1 | Did the selection process minimise researcher bias? | Assessment of citations from a second broad search was not well reported suggesting Moderate RoB, but the initial search was performed very stringently and supported by snowballing and accessing known experts, so overall we rate RoB as Low. | Low RoB |
| 3.2 | If the selection process was done in stages, were there clear criteria for stage completion | Yes | Very low RoB |
| 4 | How many (otherwise eligible) papers/studies could not be accessed | 0 | Very low RoB |

(SoF) recommended by the GRADE tool are a practical way of reporting certainty assessments.

Authors should:

- Report the overall level of certainty in the body of evidence (such as high, moderate, low, or very low) for each important outcome.
- Provide an explanation of reasons for rating down (or rating up) the certainty of evidence (such as in footnotes to an EP or SoF table). Explanations need to be easily understood by the target audience.
- Communicate certainty in the evidence wherever results are reported (that is, abstract, EP or SoF tables, results, conclusions), using a format appropriate for the section of the review.

The use of GRADE summary tables is recommended but not mandatory.

### *Examples*

Ali et al. [28] investigated the current quality of evidence concerning the benefits and limitations of AOP [Aspect-Oriented Programming] approaches compared to non-AOP approaches. They assessed strength of evidence against four criteria: Study Design, Study Quality, Consistency and Directness. For study design they say:

> *With respect to study design, the majority of the primary studies were observational. Only seven (31.8%) primary studies are experiments (see Section 4.1). Thus, according to GRADE [34], our initial categorization of the total evidence in this review from the perspective of study design is low.*
>
> Quote 28: Study Design, Ali et al [28, p. 882]

For study quality they point out that issues of bias, validity and limitations "were poorly addressed", with only three studies considering experimenter bias and only four explicitly discussing limitations while 50% of studies did not discuss limitations explicitly. Thus, they concluded that from the viewpoint of study quality, quality of evidence was low.

For consistency they report inconsistent results for the effect on code size, modularity, changeability, understandability, maintainability, performance, and exception handling. They conclude that from the viewpoint of consistency, quality of evidence is low.

With respect to directness they considered subject type (e.g., professional or students), which languages were used, the software systems used, the study setting and the outcome measures. They report that only four of the studies used human subjects, and only one of these used practitioners, while the other studies used graduate students or lecturers. Although none of the studies took place in an industry setting, most involved non-trivial software systems that were comparable to systems used in industry and the outcome measures were the same as those used in industrial settings. They conclude that the quality of evidence for directness was moderate to low.

Combining the four GRADE elements, Ali et al. [28, p. 883] assess the overall quality of evidence to be Low. They conclude that estimates of effect size are unreliable and further research is necessary.

Continuing the running example of the comparison between single company and cross-company model:

> *Our investigation resulted in three findings:*
>
> F1 *Best fitting models based on single company data are significantly more accurate than best fitting models based on cross-company data using comparisons based on the MdMRE metric ($p \leq 0.039$).*
>
> F2 *The best fitting cross-company modelling method will not necessarily correspond to the best fitting single company modelling method.*
>
> F3 *Large data sets are not necessary to construct single company estimation models as good or better than cross company models.*
>
> *Table 11 shows our assessment of the quality of the evidence supporting our individual findings. We discuss the processes of assess the quality of the body of evidence for each finding below.*

TABLE 11: Assessment of Certainty in the Body of Evidence for Three Findings (see Running Example 19)

| Issue | Comment | Assessment |
|---|---|---|
| F1 | Models based on cross-company data are less accurate than the models based on a company's own data | |
| F1 Primary Study RoB | The accuracy metric MdMRE is biased which impacts comparisons of estimation methods. The studies were not randomized trials. However, the setting of the studies were realistic and there did not appear to be any bias in the selection of single company data sets. | Moderate RoB |
| F1 Missing Data | Weakness wrt selection process for a second search (with date restrictions removed), but the first search process was very rigorous and included snowballing & contacting experts. In addition, primary studies addressed all main cross-company data sets. | Low RoB |
| F1 Imprecision | The length of the 95% confidence interval for the probability that a cross-company estimation was significantly worse than a model constructed from single company data was $0.48 - 0.02 = 0.46$. This close to the maximum possible value the interval could take and still be statistically significant, suggesting the estimate of the probability is very imprecise. | High Imprecision |
| F1 Inconsistency | Initial inconsistencies in results were substantially reduced by comparing the best cross-company estimate with the best single company models. Sensitivity analysis suggested that the finding is relatively robust. | Low Inconsistency |
| F1 Indirectness | Only one primary study used a single company data set completely independent of the cross-company data set | High Indirectness |
| F1 | Overall Quality of Evidence | Very Low Quality |
| F2 | The best cross-company estimation models may not be the same as the best single company models | |
| F2 Methodology | There is no reason to believe the effects are artifacts of the data sets or the estimation methods. Furthermore, since some studies only used one or two methods, the results may underestimate the extent of the phenomenon | Minor concerns |
| F2 Coherence | The effect was observed across different data sets, estimation methods, and sample sizes | Minor concerns |
| F2 Data Adequacy | The cross-company data sets used in the studies are the same data sets that are proposed to assist building single company estimation models. The single company data sets included both large and small data sets. | Minor concerns |
| F2 Relevance | The issue is central to the proposition that cross-company data sets can be used for single company estimates | No concerns |
| F2 | Overall Quality of Evidence | Low Quality |
| F3 | Large data sets are not necessary to construct single company estimation models as good or better than cross company models. | |
| F3 Methodology | The studies were not randomized controlled trials but the data sets contain industry software project data from multiple companies. These data sets would be the source of cross-company project data for companies trying to improve their cost estimation models. | Minor concerns |
| F3 Coherence | Eight of 10 studies with small single company data sets reported single company models that out-performed the cross-company models. Also the two primary studies with largest single company data sets reported the worst single company model accuracy. Thus, 10 of the 12 sources of evidence supported the finding. | Minor Concerns |
| F3 Data Adequacy | The cross-company data sets used in the studies are the same data sets that are proposed to assist building single company estimation models. The single company data sets included two large and a 10 small data sets. | Minor concerns |
| F3 Relevance | This issue is extremely important for companies assessing whether it is better to invest in data collection or access to cross-company data. | No concerns |
| F3 | Overall Quality of Evidence | Moderate Quality |

### *Finding F1: Single Company Predictions More Accurate than Cross-company Predictions*

*F1 is based on assessing the MdMRE difference between single company model predictions and cross-company model predictions obtained from seven independent primary studies reporting a total of 12 independent single company data sets and five different cross-company data sets.*

*Our detailed assessment of the risk of bias due to methodological weakness in the primary studies is shown in Table 7. Our detailed assessment of the risk of bias due to missing data (in particular, missing primary studies) is shown in Table 10. In this section, we integrate these results with an assessment of Indirectness and Inconsistency.*

*For* Inconsistency, *our sensitivity analyses suggested our results were reasonably robust. We therefore assess the level of Inconsistency to be Low.*

*For* Imprecision, *the length of the confidence interval of the probability that a cross-company model would outperform, or be as good as, a single company model suggested that our estimate of the probability is very imprecise. Thus, we assess the level of Imprecision to be Moderate.*

*In the case of* Indirectness, *all but one of the primary studies (i.e. Study S4) involved single company data set that was already held in the cross-company data set.*

*In addition, all data sets except the Tukutuku data set used function points as a size metric. It is unclear whether the function point values were post delivery measures or estimates made during the requirements specification. If post delivery measures were used, then the accuracy of the estimation models may be inflated (because they would be unaffected by inconsistencies between function point measures assessed during requirements and function point measured post delivery).*

*Overall the set of studies provided very little information as to how the a single company would actually use cross-company data to support their own cost estimation processes. Thus, we assess that the level of Indirectness is* High.

*Given the problems associated with Indirectness, we conclude that the overall quality of the evidence relating to the estimate of average MdMRE is Very Low meaning that "Any estimate of effect is very uncertain".*

### *F2 The Best Cross-Company Estimation Method Can Differ from the Best Single Company Estimation Method*

*There were five examples where the best fitting single company estimation method was not the same as the best fitting cross-company estimation method. In one case, the MdMRE difference was unchanged, in three cases the direction of the MdMRE value changed direction sufficiently to favour single company cost model, and in the final cases the MdMRE value difference was reduced from -10 to -7. This finding identifies that a possible additional risk to using a cross-company model with little or no single company data is that the cross-company model may be less accurate than expected when used for single-company predictions.*

*This finding is not the outcome of a statistical synthesis, so we assess its quality from the viewpoint of a qualitative finding using the CERQual assessment process [54]:*

*In terms of Methodological Limitations, although the primary studies are not randomized controlled trials, the data sets used in studies are the source of cross-company project data that would used by single companies. Thus, the data used by the studies was completely realistic. In addition, the current set of studies may have* underestimated *the prevalence of F2 because some studies only used one or two cost estimation modelling methods so had little chance of detecting differences between the accuracy of different cost estimation modelling methods. For example, the studies of the Tukutuku dataset used only step-wise regression after applying the logarithmic transformation to size and effort variables). We conclude that there are Minor concerns about Methodological Limitations.*

*In terms of Coherence, the phenomenon was observed for four of the five different cross-company data sets, for both large and small single company data sets, and for a variety of different cost estimation methods. Thus, the evidence supports the view that F2 is a reasonably wide-spread phenomenon. Given the available cross-company data sets, it would unlikely that further research would change these observations. We conclude that there are Minor concerns about Coherence.*

*In terms of Data Adequacy, the cross-company data sets used in the studies are the same ones that are proposed to assist single companies construct estimation models. The single company data sets include examples of both large and small data sets. There is no reason to believe that this result is an artefact of the data sets or analysis method. However, only five of the studies reported the phenomenon. Thus, we assess that there are Moderate concerns about Data adequacy.*

*In terms of Relevance, F2 is extremely important from the viewpoint of assessing the value of cross-company data to support single company cost estimation. We conclude that there are No concerns about Relevance.*

*Overall, we rate confidence in the quality of the evidence supporting F2 to be Low meaning that meaning that "It is possible that the review finding is a reasonable representation of the phenomenon of interest" [54].*

### *F3 Large Single Company Data sets are not Essential for Reasonable Estimation Models*

*This finding is not the outcome of a statistical synthesis, so we assess its quality from the viewpoint of a qualitative finding using the CERQual assessment process [54].*

*As discussed previously the studies do not report randomized trials, but do report analyses of relevant data sets. In the context of sample size, the selection of single company data sets was based on the available data, and there was no indication of any systematic bias in their selection which would cause us to doubt the results. With respect to data analysis the magnitude relative error is a biased metric, but the median value is more reliable than the mean and the bias is mainly a problem for studies comparing estimation methods. Overall we assess there to be Minor concerns about* Methodological Limitations.

*In terms of Data Adequacy, we have 10 small single company data sets compared with two large single company data sets, so we have more evidence about small data sets. However, the two large data sets poor accuracy, which also support F3. Only one of the single company data sets was completely independent of the cross-company data set. We conclude there are Minor concerns about* Data Adequacy.

*In terms of Coherence, we observed that eight of the 10 single company data sets with less than 20 projects had MdMRE levels of less than 30%. In contrast, the two largest single company data sets had the two worst accuracy levels (MdMRE values of 39% and 60%). Overall our observations are consistent with F3 and there are Minor concerns about* Coherence.

*Companies considering the use of cross-company data sets to assist effort estimation need to assess whether they have sufficient data to be confidence in building estimation models based on their own data. We conclude that there are No concerns about* Relevance.

*Overall, we rate confidence in the quality of the evidence supporting F3 to be Moderate meaning that "It is likely that the review finding is a reasonable representation of the phenomenon of interest" [54].*

Example 19: Quality of Evidence—Running Example Continued

# 5 Discussion and Limitations

## 5.1 Discussion - General Issues

A discussion section is required for all secondary studies, although some of the sub-items are irrelevant for mapping studies. Authors should avoid simply repeating the outcomes of any analysis and synthesis, but should discuss the results in terms of the practical implications of the findings in the context of existing knowledge and any limitations that should be place on the review findings.

## 5.2 Item 23a Relationship with Other Evidence

PRISMA 2020 Definition: Authors should provide a general interpretation of the results in the context of other evidence.

For mapping studies authors should compare their results with other secondary studies that address either the same or a related issue.

### *Explanation*

Discussing other evidence should help readers interpret the findings. For example, authors might:

- Compare the current results to results of other similar systematic reviews (such as reviews that addressed the same question using different methods or that addressed slightly different questions) and explore possible reasons for any disagreements.
- Summarize other additional information not explored in the review, such as the training costs associated with new working methods, or surveys gauging the attitudes and preferences of software practitioners.

### *Examples*

In their tertiary study of SE systematic reviews which updated two previous studies, da Silva et al. [29] compare their results with the previous tertiary studies as follows:

> *Our study shows three important changes in the study set from the previous tertiary studies [18,19]. First, the coverage of topics in software engineering increased, and the concentration in a few topics decreased. Second, the number of researchers and, consequently, organisations undertaking systematic reviews increased and became more globally distributed. Finally, we found proportionally more mapping studies than conventional systematic reviews in our study.*
>
> Quote 29: Tertiary Study Comparisons [29, p. 911]

In his SR investigating the accuracy of effort estimates obtained by expert opinion compared with those from formal models, Jørgensen [27] concluded:

> *If, as suggested in MacDonell and Shepperd (2003), there is a high degree of independence between estimates based on common effort estimation models and expert judgement, and it is difficult to devise rules for selecting the most accurate estimation method, the solution seems to be to use a combination of models and experts.*
>
> Quote 30: Discussion of Cost Estimation Study Results [27, p. 460]

Continuing the running example of the comparison between single company and cross-company models:

> *One unexpected result we found among the primary studies was that the two largest single companies (i.e., Study S6a and Study 10) had the worst average MdMRE values for the single company estimation models (60 and 39 for S10 and S6a respectively). Thus, although the project came from a single company, it is possible that the projects actually came from very different types of software projects. The COCOMO [63] and DMR data set [64] are well-known examples of multi-type single company data sets. Appropriate analysis of such a data set relies on a nominal scale attribute that defines the different projects types and allows models to be developed for each type. Failing to recognize important subsets in a large single company set data will reduce the accuracy of any model built using the data.*
>
> Example 20: Relationship With Other Evidence—Running Example Continued

## 5.3 Item 23b Evidence Limitations

PRISMA 2020 Definition: Authors should discuss any limitations of the evidence included in the review.

This item is irrelevant for mapping studies.

### *Explanation*

Discussing the completeness, relevance, 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 many studies with a small number of participants, leading to imprecise estimates;
- have concerns about risk of bias in studies or missing results;
- have included studies that only partially or indirectly address the review question, such as finding only laboratory studies with student participants, leading to concerns about their relevance to practitioners.

The assessments of certainty (or confidence) in the body of evidence (item 22) can support the discussion of such limitations.

### *Examples*

Summarising the limitations of his study of perspective-based reading, Ciolkowski [26] identified the following three issues that might impact the certainty of the evidence:

1) Few studies that investigated design or code inspection.
2) There were a large number of confounding variables.
3) The studies only included formal experiments, i.e., there were no industry case studies.

Specifically, on the topic of confounding variables he says:

> High number of confounding variables*: Many potential moderator variables exist, and none of them explains the existing heterogeneity satisfactorily. We still have to find a better way to deal with finding the best combination of context variables. In our case, we used exploratory cluster and factor analyses.*
>
> Quote 31: Limitation of a PBR Meta-Analysis [26, p. 143]

## 5.4 Item 23c Review Process Limitations

PRISMA 2020 Definition: Authors should discuss any limitations of the review processes used.

*Explanation*

To help readers understand the trustworthiness of the review findings, authors need to discuss review process limitations, avoidable or unavoidable. For example, authors might acknowledge adopting processes that:

- Risk missing primary studies, such as the decision to restrict eligibility to studies in English only, searching only a small number of databases, excluding technical reports and white papers.
- Risk introducing experimenter bias, such as by having only one reviewer screen database records or extract data,
- Risk of introducing data bias by not contacting study authors to clarify unclear information, or being unable to access all potentially eligible study reports.
- Risk conclusion bias by having insufficient data to carry out some planned analyses. For example, if many primary studies report one outcome value (e.g., developer effort), but too few measure the correctness or quality of the task outcome for a trustworthy statistical analysis, there is a risk that a method that requires less effort may be identified as preferable even though it may in fact result in the production of task deliverables of lower quality.

Authors should discuss the potential impact of the limitations. Some limitations may affect the validity of the review findings, others may not. For example, if the search process did not find any foreign language papers, restriction to English language papers did not cause any actual problem.

For mapping studies that adopted the standard systematic review process for search, selection and data extracting, this item will be irrelevant.

*Examples*

Kitchenham and Brereton [9] identified poor initial agreement achieved on study quality as the main limitation of their systematic review process. They concluded that their assessment of study quality might be rather error prone and explain how they addressed that issue as follows:

> *To address this we have reported not just the quality score but our assessment of the type of validation performed and the context of the validation, which provide some additional indication of the stringency of the validation exercise.*
>
> Quote 32: Main Limitation Reported in an SR on SR Process Research [9, p. 2069]

They also discuss the limitation due to using the extractor-checker method.

### 5.5 Item 23d Implications for Practice, Policy and Future Research

PRISMA 2020 Definition: Authors should discuss implications of the results for practice, policy, and future research.

Mapping studies should concentrate on discussing future research.

*Explanation*

There are many possible end users of systematic review evidence such as practitioners, researchers, managers, and educators, each of whom would like to know what actions they should take given the findings of a review.

Authors should:

- discuss the implications of the research for practice, policy and education,
- make explicit recommendations for future research, not just general comments such as "more research is needed".

*Examples*

Kitchenham and Brereton [9] identify eleven implications of their review on the current software engineering SR guidelines, the first of which being:

> *To remove the proposal for constructing structured questions and using them to construct search strings. It does not work for mapping studies and appears to be of limited value to SRs in general since it leads to very complex search strings that need to be adapted for each digital library.*
>
> Quote 33: SR Guidelines Changes [9, p. 2068]

Kitchenham and Brereton [9] also suggests three areas for further research:

- SR Process tools.
- Large-scale evaluations of textual analysis tools.
- Quality evaluation of SE papers.

Continuing the running example of the comparison between single company and cross-company model:

> *Overall none of our findings supported the idea that cross-company data and be used to assist single company estimation. In particular, there was no standard cost estimation method that could be guaranteed to construct the best model for a single company, so it is unclear how to build an appropriate single company model from cross-company data. Furthermore, by the time a single company has sufficient data to decide which cross-company model best fits its data, it would have sufficient data to produce its own model, which would probably be more accurate than any cross-company model.*
>
> *We therefore recommend that companies concentrate on collecting data about their own software projects and building estimation models from their own data.*
>
> *Further research on this topic should consider the mechanics of how existing cross-company estimation data could be incorporated into single company estimation processes. For example, Bayesian approaches might be a means of integrating single company data (including expert-opinion based estimating processes) with cross-company estimation models and data.*
>
> Example 21: Recommendations for Practice and Research—Running Example Continued

## 6 Issues Related to Scientific Ethics

This section covers issues related to scientific ethics, including access to the protocol, deviations from the protocol, and research data availability, and conflicts of interest.

### 6.1 Protocol Registration Information

*Explanation*

Readers should be able to compare the initial pre-specified plan for the SR with the final SR report. This allows readers to assess whether any deviations from the plan might have introduced bias.

### 6.2 Item 24a Registration Information

PRISMA 2020 Definition: Authors should provide registration information for the review, including register name and registration number, or state that the review was not registered.

*Explanation*

SE protocols are not usually registered, but if a protocol has been registered, authors should report the registration information.

### 6.3 Item 24b Protocol Access

PRISMA 2020 Definition: Authors should indicate where the review protocol can be accessed, or state that a protocol was not prepared.

*Explanation*

All SRs should produce a protocol. It is good scientific practice and can reduce the length of the methods section by providing a source of detailed information about search, selection, and data extraction processes. It is important to ensure that the protocol is accessible and will remain accessible. Practical problems arise if the protocol can only be obtained from members of the research team while they are employed at a specific institution.

Authors need to specify the link to the protocol, or explain why no protocol is available. In seeking a long-term, reliable place for your protocol, you may consider services such as OSF[5] (a free, open platform to support your research and enable collaboration that allows users to create project folders, pre-register study protocols, and even store data and code files for public access) or Zenodo[6] (a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN, but open to researchers from outside the EU).

*Examples*

Beecham et al [18] state in the introduction to their Method section, after itemising the steps they undertook in their SR, that:

> *These steps are detailed in our protocol (See [3] or http://homepages.feis.herts.ac.uk/~ssrg/MOMSEProto.htm).*
>
> Quote 34: Protocol Access, [18, p. 862]

### 6.4 Item 24c Deviations from the Protocol

PRISMA 2020 Definition: Authors should describe and explain any amendments to information provided at registration or in the protocol.

5. https://osf.io/
6. https://zenodo.org/

*Explanation*

It is difficult to anticipate all scenarios that will arise, necessitating some clarifications, modifications, and changes to the protocol (e.g., the available data may not be amenable to the planned meta-analysis). For reasons of transparency, authors should report details of any amendments. Amendments could be recorded in various places, including the full text of the review, a supplementary file, or as amendments to the published protocol or registration record.

*Examples*

Kitchenham et al. [11] mention a major deviation from their SR protocol on meta-analysis methods used in SE families of experiments as follows:

> *The major deviation from the protocol and the results reported in this paper is that originally we had assumed it would be appropriate to concentrate on reproducibility, but as our investigation progressed we realized that we needed to consider the reasons for lack of reproducibility, that is, consider in more detail the validity of the meta-analysis process. Furthermore, validity is the key issue, because it is not useful to reproduce an invalid result.*
>
> Quote 35: Deviations from Protocol of SR on Meta-Analysis Methods [11, p. 356]

### 6.5 item 25 Support

PRISMA 2020 Definition: Authors should describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.

*Explanation*

Authors of any research report need to be transparent about the sources of support received to perform the research (either direct monetary support to cover staff or tool costs) or provision of staff from the funding organisation to help with the review process. There is potential for bias in the review findings arising from such involvement, particularly when the funder or sponsor has an interest in obtaining a particular result.

Authors need to :

- Describe sources of financial or non-financial support for the review, specifying relevant grant ID numbers for each funder. If no specific financial or non-financial support was received, this should be stated.
- Describe the role of the funders or sponsors (or both) in the review. If the sponsors/funders had no role in the review, this should be explicitly confirmed, for example, by stating "The funders had no role in the design of the review, data collection and analysis, decision to publish, or preparation of the manuscript."

*Examples*

In their paper on the reproducibility of code smells research, Lewowski and Madeyski [17] report

> *This research was partly financed by Polish National Centre for Research and Development, Poland grant POIR.01.01.01-00-0792/16: "Codebeat - wykorzystanie sztucznej inteligencji w statycznej analizie jakości oprogramowania."*

Quote 36: Declaration of Support [17, p. 13]

### 6.6 Item 26 Competing Interests

PRISMA 2020 Definition: Authors should declare any competing interests of review authors.

#### *Explanation*

If authors have relationships with organizations that have an interest in the review outcomes (e.g., serving as a consultant or running paid training courses), such relationships can negatively impact the credibility of results. Authors need to:

- Disclose any authors' relationships or activities that readers could consider pertinent or could have influenced the review.
- If any competing interests are declared, report how they were managed for the specific review.

#### *Examples*

In their paper on the reproducibility of code smells research, Lewowski and Madeyski [17] report:

> *Declaration of competing interest*
> *No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.infsof.2021.106783.*
>
> Quote 37: Declaration of No Competing Interest [17, p. 13]

### 6.7 Item 27 Availability Of Data, Code and Other Materials

PRISMA 2020 Definition: Authors should 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.

#### *Explanation*

Sharing of data, analytic code, and other materials enables others to reuse the data, check the data for errors, attempt to reproduce the findings, and understand more about the analysis than may be provided by descriptions of methods [65].

Authors should report

- Which of the following are publicly available: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.
- Report which of the following are publicly available: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.
- If data, analytic code, or other materials will be made available upon request, provide the contact details of the author responsible for sharing the materials and describe the circumstances under which such materials will be shared.

#### *Examples*

In their paper on the reproducibility of code smells research, Lewowski and Madeyski [17] report:

> *Supplementary material related to this article can be found online at https://doi.org/10.1016/j.infsof.2021.106783.*
>
> Quote 38: Accessibility to Additional Material [17, p. 13]

## 7 Reflections on the Examples

We have identified examples from both quantitative and qualitative systematic reviews. We have not concerned ourselves with mapping studies, since they are much simpler to report than other types of systematic secondary study when using the SEGRESS items.

We included a running example covering the reporting process for risk of bias, risk of missing data and the certainty of the body of evidence because it is important to see how the SEGRESS items interact with each other to provide the overall assessment of the quality of the body of evidence. Another benefit of our running example is that it provides an example of how GRADE and GRADE-CERQual evaluations can differ for different findings. Applying certainty assessment to multiple findings has not been reported in any published SE systematic review.

The process of assigning risk of bias assessments to individual primary studies, the set of primary studies as a whole and other elements such as Indirectness and Inconsistency is subjective. To have any level of trust in the results, assessments should be based on having several researchers undertake the process independently. For the purposes of our example, however,the assessment was done only by Kitchenham because she was one of the authors of the SR on which the example was based (see [8], [6], [7]).

In our opinion, the most problematic aspect of assessing risk of bias is assessing the risk of missing data/projects. Other issues depend on the decisions made by the authors of the primary studies, but the risk of missing data depends on the authors of the SR. Checklists of criteria that can be used to assess the rigour of the search process can help, but the authors of the SR are also the ones who choose or develop the checklists. The best advise we can give is for authors to ensure that they report their search and selection process as clearly and completely as possible, so that readers of the SR can judge the accuracy of the risk of missing data for themselves.

In terms of using the SEGRESS items, we found 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. One issue we found problematic was the ordering of items 13 and 20. Specifically, the ordering of the PRISMA items seems to imply reporting sensitivity analyses after reporting heterogeneity investigations. However, it is better to make sure that you have a robust synthesis before investigating whether any factors can explain inconsistencies. We, therefore, suggest that users of SEGRESS address items 13e and 13f and items 20c and 20d in the order that is most appropriate for their specific SR report.

## REFERENCES


[1] 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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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

[9] 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.

[10] B. Kitchenham and P. Brereton, "Protocol for a Mapping Study of SLR Process Research in Software Engineering," 2012, version 5.2. [Online]. Available: https://madeyski.e-informatyka.pl/download/Kitchenham12Protocol.pdf

[11] B. Kitchenham, L. Madeyski, and P. Brereton, "Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment," *Empirical Software Engineering*, vol. 25, no. 1, pp. 353–401, 2020. [Online]. Available: https://doi.org/10.1007/s10664-019-09747-0

[12] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – a systematic literature review," *Information and Software Technology*, vol. 51, pp. 7–15, 2009.

[13] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Protocol for a tertiary study of systematic literature reviews and evidence-based guidelines in it and software engineering," 2007, version 3. [Online]. Available: https://madeyski.e-informatyka.pl/download/Kitchenham07Protocol.pdf

[14] B. Kitchenham, R. Pretorius, D. Budgen, P. Brereton, M. Turner, M. Niazi, and S. Linkman, "Systematic literature reviews in software engineering — a tertiary study," *Information & Software Technology*, vol. 52, pp. 792–805, 2010.

[15] B. A. Kitchenham, O. P. Brereton, and D. Budgen, "Protocol for Extending an existing Tertiary study of Systematic Literature Reviews in Software Engineering," 2008, version 3. [Online]. Available: https://madeyski.e-informatyka.pl/download/Kitchenham08Protocol.pdf

[16] T. Lewowski and L. Madeyski, *Code Smells Detection Using Artificial Intelligence Techniques: A Business-Driven Systematic Review*. Cham: Springer International Publishing, 2022, pp. 285–319. [Online]. Available: https://doi.org/10.1007/978-3-030-77916-0_12

[17] T. Lewowski and L. Madeyski, "How far are we from reproducible research on code smell detection? a systematic literature review," *Information and Software Technology*, vol. 144, p. 106783, 2022. [Online]. Available: https://doi.org/10.1016/j.infsof.2021.106783

[18] 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.

[19] 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.

[20] 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

[21] 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.

[22] F. Q. d. Silva, S. S. Cruz, T. B. Gouveia, and L. F. Capretz, "Using meta-ethnography to synthesize research: A worked example of the relations between personality and software team processes," in *2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement*, 2013, pp. 153–162.

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

[24] 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.

[25] 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.

[26] 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

[27] 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.

[28] 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.

[29] F. Q. da Silva, A. L. Santosa, S. Soares, A. C. C. França, C. V. Monteiro, and F. F. Maciel, "Six years of systematic literature reviews in software engineering: An updated tertiary study," *Information & Software Technology*, vol. 53, no. 9, pp. 899–913, 2011.

[30] F. Elberzhager, A. Rosbach, J. Münch, and R. Eschbach, "Reducing test effort: A systematic mapping study on existing approaches," *Information and Software Technology*, vol. 54, pp. 1092–1106, 2012.

[31] 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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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

[37] M. Sośnicki and L. Madeyski, "ASH: A New Tool for Automated and Full-Text Search in Systematic Literature Reviews," in *Computational Science – ICCS 2021*, M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. Sloot, Eds. Cham: Springer International Publishing, 2021, pp. 362–369.

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

[39] 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

[40] 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

[41] 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.

[42] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg, "Systematic review: A systematic review of effect size in software engineering experiments," *Information & Software Technology*, vol. 49, no. 11-12, pp. 1073–1086, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2007.02.015

[43] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, "Robust Statistical Methods for Empirical Software Engineering," *Empirical Software Engineering*, vol. 22, no. 2, pp. 579–630, 2017. [Online]. Available: http://link.springer.com/content/pdf/10.1007%2Fs10664-016-9437-5.pdf

[44] D. Bowes, T. Hall, and D. Gray, "Dconfusion: a technique to allow cross study performance evaluation of fault prediction studies," *Automated Software Engineering*, vol. 21, no. 2, pp. 287–313, 2014.

[45] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE'08. British Computer Society, 2008, pp. 68–77.

[46] 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

[47] Y. Eaves, "A synthesis technique for grounded theory data analysis," *Journal of Advanced Nursing*, vol. 35, no. 5, pp. 654–663, 2001.

[48] 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.

[49] N. Berkman, K. Lohr, and M. A. et al., *Chapter 15 Appendix A: A Tool for Evaluating the Risk of Reporting Bias (in Chapter 15: Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update). Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007- 10056-I)*. Rockville, MD: Agency for Healthcare Research and Quality, 2013, no. 13(14)-EHC130-EF.

[50] N. Meader, K. King, A. Llewellyn, G. Norman, J. Brown, M.Rodgers, T. Moe-Byrne, J. Higgins, A. Snowden, and G. Stuart, "A checklist designed to aid consistency and reproducibility of GRADE assessments: development and pilot validation," *Systematic Reviews*, vol. 3, no. 82, 2014.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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

[56] F. Selleri Silva, F. S. F. Soares, A. L. Peres, I. M. de Azevedo, A. P. L. Vasconcelos, F. K. Kamei, and S. R. de Lemos Meira, "Using cmmi together with agile software development: A systematic review," *Information and Software Technology*, vol. 58, pp. 20–43, 2015.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] T. Dybå, D. I. Sjøberg, and D. S. Cruzes, "What works for whom, where, when, and why?: On the role of context in empirical software engineering," in *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '12. New York, NY, USA: ACM, 2012, pp. 19–28.

[62] 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.

[63] B. W. Boehm, *Software Engineering Economics*. Prentice-Hall, 1981.

[64] J.-M. Desharnais, "Analyse statistique de la productivité des projets de développment en informatique à partir de la technique des points de fonction," Master's thesis, Université du Québec à Montréal, 1988.

[65] 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