e-Informatica Software Engineering Journal, Volume 0, Issue 0, 0000

A Quality Assessment Instrument for Systematic Literature Reviews in Software Engineering

Muhammad Usman*, Nauman Bin Ali*, Claes Wohlin*
*Department of Software Engineering, Blekinge Institute of Technology, Sweden.
muhammad.usman@bth.se, nauman.ali@bth.se, claes.wohlin@bth.se

Abstract

Background: Systematic literature reviews (SLRs) have become a standard practice as part of software engineering (SE) research, although their quality varies. To build on the reviews, both for future research and industry practice, they need to be of high quality. **Aim:** To assess the quality of SLRs in SE, we put forward an appraisal instrument for SLRs. **Method:** A well-established appraisal instrument from research in healthcare was used as a starting point to develop the instrument. It is adapted to SE using guidelines, checklists, and experiences from SE. The first version was reviewed by four external experts on SLRs in SE and updated based on their feedback. To demonstrate its use, the updated version was also used by the authors to assess a sample of six selected systematic literature studies. **Results:** The outcome of the research is an appraisal instrument for quality assessment of SLRs in SE. The instrument includes 15 items with different options to capture the quality. The instrument also supports consolidating the items into groups, which are then used to assess the overall quality of an SLR. **Conclusion:** The presented instrument may be helpful support for an appraiser in assessing the quality of SLRs in SE.

1. Introduction

To establish evidence-based practices in software engineering (SE), Kitchenham [27] proposed the use of systematic literature reviews (SLRs) to identify, appraise and synthesise evidence reported in the scientific literature. Today the method is well-accepted in SE. This is illustrated by the growing number of published SLRs. The number of SLRs has increased rapidly since the introduction of the guidelines [4].

The reliability of conducting SLRs as a method needs to be maintained, ensuring trust in the results. Several researchers have proposed guidelines and checklists to design, conduct and report SLRs. However, relatively little work has been done on the critical appraisal of the SLRs. Several SE researchers have used an interpretation of the criteria used by the Centre for Reviews and Dissemination (CRD) at the University of York to include an SLR in their Database of Abstracts of Reviews of Effects (DARE) [11]. However, these questions are insufficient to reveal significant limitations in SLRs [8, 4, 13]. These limitations act as a motivation for the research presented here.

In 2007, Kitchenham and Charters [29] introduced the DARE criteria [11] for quality assessment of SLRs in SE. The criteria come from healthcare and medicine. DARE included four criteria/questions. Since then, DARE has been updated to include five criteria/questions. According to Costal et al. [13], DARE is the most used quality assessment instrument of SLRs in SE, although many customize the criteria due to missing aspects in DARE or adapting it for their specific study.

According to Shea et al. [42]: "*The Cochrane Collaboration Handbook provides a comprehensive guide for review authors, but it does not provide a concise critical appraisal instrument for completed reviews.*" Thus, according to Shea et al. [42] the DARE criteria are primarily for authors of SLRs and not for assessing the quality of published reviews. According to CRD [22], the five criteria are used to decide whether to include an SLR in their database or not, which may be why it is not perceived to be sufficiently "concise and critical". It is sufficient that four out of five criteria are met for the SLR to be included in the CRD database. Thus, when combining the viewpoints of DARE, it implies that authors should ensure that they cover the DARE criteria for their SLR being accepted into the CRD database, i.e. the criteria are primarily inclusion/exclusion criteria. Thus, as noted, it is a quite high-level assessment, and it is not aimed at scrutinizing SLRs from a quality perspective. Despite the main objective of DARE, it has been used to assess the quality of published SLRs in SE. In a similar way as DARE is used in relation to the CRD database, DARE may be used as a screening instrument when deciding to include or exclude an SLR in a tertiary study. However, it is not suitable as a quality assessment instrument of SLRs. Costal et al. [13] conclude that there is a need for a comprehensive framework covering more aspects of quality than DARE.

To address quality assessment of published SLRs of randomized trials, AMSTAR was introduced in 2007 in medicine and healthcare. AMSTAR is, according to Shea et al. [42], one of the most used instruments for quality assessment. Through extensive use and validation of AMSTAR in healthcare and medicine [40], several necessary improvements were identified [9, 46]. AMSTAR 2 [42] was developed to address the limitations of AMSTAR, like handling non-randomized trials and being aligned with the revised Cochrane risk of bias instrument, which is described by Sterne et al. [43].

With the development of AMSTAR and its further improvement through AMSTAR 2, it is time to upgrade the quality assessment of published SLRs in SE instead of using the DARE criteria. Thus, we concur with Costal et al. [13] concerning the need for a more comprehensive framework covering more quality aspects, and on a more detailed level. We should continue to learn from other disciplines that are conducting SLRs. Therefore, building on the success of AMSTAR and its successor AMSTAR 2, we propose an adaptation of AMSTAR 2 to SE, which we call QAISER (**Q**uality **A**ssessment **I**nstrument for **S**oftware **E**ngineering systematic literature **R**eviews).

Our approach when developing QAISER has several salient features focusing on increasing the reliability of the research outcome. We based our work on a well-accepted and validated instrument as a foundation (i.e., AMSTAR 2) [42]. To ensure an appropriate adaptation to SE, we collected and relied on a comprehensive set of documents with guidelines and best practices for conducting SLRs in SE. We followed a systematic and well-documented process to develop QAISER with several internal validation steps involving multiple researchers. Furthermore, we invited some leading experts in evidence-based SE research to conduct an external validation of QAISER. We also demonstrate the applicability of QAISER by using it to assess a sample of six selected systematic literature studies. In each step of the process, QAISER was updated based on the feedback received and internal discussions. The outcome of the process, i.e., QAISER, is the main contribution of the paper.

Our main objective is to support appraisers in assessing quality of completed SLRs, we believe that authors of SLRs may also use QAISER to help them with improving the quality of their SLR before submitting the research for assessment.

The remainder of the paper is organised as follows. Section 2 presents an overview of the main critical appraisal instruments used in both SE and evidence-based medicine. Section 3 describes in detail the method undertaken for developing QAISER. In Section 4, it is described how QAISER evolved into the latest version, which is the main outcome of the research presented. Section 5 describes how QAISER can be used to assess SLRs. In Section 6, we reflect on the reliability

of QAISER. Section 7 describes the guidance document and shares our reflections about applying QAISER on six example SLRs. The threats to validity are presented in Section 8. Section 9 discusses the implication of the results. Section 10 concludes the paper, presents future research directions and our ambition to support broader adoption of QAISER. Finally, the QAISER instrument is provided in Appendix A (attached as supplemental material), and a guidance document supporting the instrument can be found in Appendix B (also attached as supplemental material).

2. Related work

A prerequisite for a quality appraisal is that we pose the right questions. In the first version of the guidelines for systematic literature reviews in SE, Kitchenham [27] identified two sets of questions from Greenhalgh [21] and Khan et al. [25] to review any existing SLRs on a topic of interest. In the 2007 update [29], Kitchenham and Charters added the CRD DARE set of four questions to the list [11]. Kitchenham and Charters [29] also applied the criteria to SLRs published between 2004 and 2007.

The proposal from Greenhalgh is very general; Khan et al.'s proposal is the most comprehensive, while the DARE criteria are brief and "simple" [29]. Among these three sets of questions proposed in the guidelines, only the DARE criteria have been widely used in the SE literature.

Kitchenham et al. [30] provided guidance to answer four of the five questions in the DARE criteria. Cruzes and Dybå [14] observed that one of the critical questions regarding synthesis had not been included in the SE guidelines for conducting SLRs and has not been used when evaluating the quality of SLRs in SE. It should be noted that the number of questions in DARE has varied over the years; it has included either four or five questions depending on the version of DARE.

Some others have developed their own interpretation of the DARE questions [36, 3]. One shared limitation of these is the lack of traceability between the proposals and the evidence/best practices used to motivate them.

Other researchers have also been concerned with assessing quality in SLRs in SE, Dybå and Dingsøyr [19] reviewed several proposals from evidence-based medicine to assess the quality of SLRs. They concluded that the MOOSE statement [44] is a very relevant reporting checklist for SLRs in SE. The MOOSE checklist has six main reporting items including 'background', 'search strategy', 'method', 'results', 'discussion' and 'conclusions'. Each item further lists actions and details that should be provided in an SLR.

In a previous study [5], we reviewed the proposals for quality assessment for SLRs both from SE and other fields. We concluded that in the SE literature, there is an awareness of reporting checklists like MOOSE, QUOROM, and PRISMA. However, SE researchers have not yet leveraged the progress in critical appraisal tools for systematic reviews.

One essential aspect related to quality assessment is the validity threats presented by authors of SLRs. Ampatzoglou et al. [7] reviewed 100 secondary studies in SE and identified the commonly reported threats to validity and the corresponding mitigation actions. They also proposed a checklist that authors can use to design an SLR with explicit consideration for common validity threats and develop an informed plan for mitigating them. The authors state that readers can also use the checklist to assess the validity of the results of an SLR. The checklist has 22 questions grouped into three categories: study selection validity, data validity, and research validity. Furthermore, for each of the 22 questions, there are 1 to 9 sub-questions.

The checklist by Ampatzoglou et al. [7] encapsulates the current state of research regarding mitigating validity threats. Also, the checklist is a useful design tool to support the design, execution and reporting of an SLR. However, we argue that it is not a tool that enables the evaluation of

Name	Approach	Domain	Awareness in SE
Ampatzoglou et al. [7]	Checklist	SE	Yes
MOOSE [44], QUOROM, PRISMA	Checklist	Medicine	Yes (see [5, 19])
PRISMA-ScR (for mapping studies), and ENTREQ and RAMESES (for qualitative reviews)	Checklist	Medicine	Yes (see [33]
SEGRESS [33]	Checklist	SE	Yes
DARE	Checklist	Medicine	Yes (see [29, 8])
DARE interpretation in SE (see [29, 8])	Assessment Instrument	SE	Yes
AMSTAR, AMSTAR 2, ROBIS	Assessment Instrument	Medicine	Yes (see [5])

Table 1: An overview of checklists and assessment instruments for SLRs

completed SLRs, e.g. should all items in the checklist be addressed?. Even as a reporting checklist, Kitchenham et al. [33] point out the following major weaknesses in Ampatzoglou et al.'s approach and proposal: (1) they present what threats to validity are reported and not what should be reported, (2) they may have underestimated the extent of validity issues in secondary studies, and (3) they mix the threats to validity for mapping and reviews. Nevertheless, their work inspired the development of some QAISER items.

Table 1 presents an overview of various design and reporting checklists and assessment instruments for the quality assessment of systematic literature reviews.

Given the lack of an appraisal tool adapted for SE, we wanted to leverage experiences from other research fields. Through an analysis of the leading appraisal tools, including ROBIS, AMSTAR, and AMSTAR 2, we identified AMSTAR 2 (A MeaSurement Tool to Assess systematic Reviews) [42] as a candidate tool for adaptation to SE [5]. AMSTAR was developed based on a review of available rating instruments and consolidated them into 11 appraisal items. It has since been extensively used and validated. AMSTAR 2 is a revised version of the tool that takes into account the systematically collected community feedback. The major updates for AMSTAR 2 are: (1) the consideration of SLRs that may include non-randomized studies and (2) an increased focus on the risk of bias evaluation.

AMSTAR 2 provides a more comprehensive coverage of important quality aspects of an SLR that are not included in the DARE criteria that are mostly used in SE [5]. AMSTAR 2 consists of 16 appraisal items and their corresponding response options and scale. Figure 1 annotates an example of an item, response, and scale from QAISER. QAISER kept the structure from AMSTAR 2. Nine QAISER items have three scale options (Yes, No, Partial Yes), while the rest have only two scale options (Yes, No). Like AMSTAR 2, QAISER includes 'Partial Yes' in cases where it is relevant to recognise partial compliance with items. In the case of the six items with only two Yes/No options, partial compliance is not an option - i.e., for these items all response options are considered equally important. However, where possible, the alternate way to achieve a 'Yes' rating is provided (Items 12, 13, and 15).

Based on an analysis of related work it was decided to use AMSTAR 2 as a basis for proposing a quality assessment instrument tailored for SE.

3. Method

This section describes the four-step process we used to develop QAISER (see Figure 2 for an overview). In the first step, we identified aspects from the evidence-based software engineering (EBSE) literature relevant for inclusion in QAISER. In the second step, we adapted AMSTAR 2

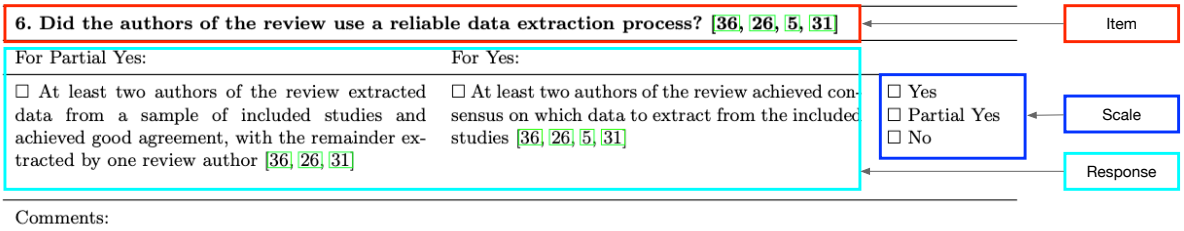

Figure 1: Items, responses and scale in QAISER.

to SE by customizing its items and responses. In the third step, we combined the outputs of the previous two steps by integrating the EBSE aspects into QAISER. In the fourth step, we validated QAISER by inviting external experts to evaluate its completeness, understandability, and relevance of its items and responses for SE. Furthermore, we also used QAISER to assess a sample of six SLRs to demonstrate its applicability.

The first two authors jointly performed Steps 1-3 of the process, while the third author - the most experienced of the three authors - independently reviewed the work. Such division of roles among the authors was introduced to have an internal continuous sanity check on the outputs of all steps. Each step is further elaborated below and the details of each step are also illustrated in Figures 3 – 6 (the bidirectional arrows in these figures indicate that an activity results in the updates to its input).

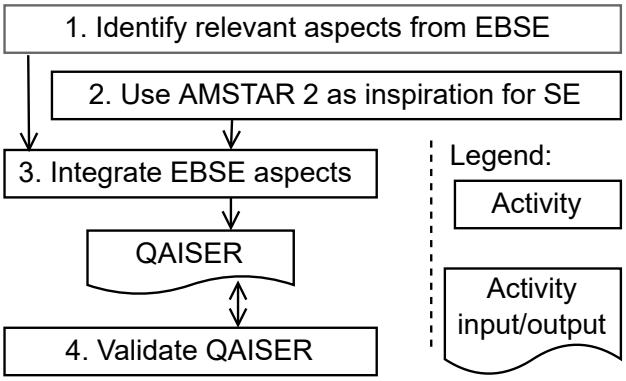

Figure 2: Overview of the QAISER development process.

Step 1: Identifying relevant aspects from the EBSE literature

In this step, we aimed to complement AMSTAR 2 with the relevant work from the EBSE literature. We followed a systematic approach to identify and analyze the relevant EBSE work (see Figure 3).

We started with analyzing a closely related and recent tertiary study on validity threats in SLRs in SE by Ampatzoglou et al. [7]. They have aggregated validity threats and corresponding mitigating actions in the form of a checklist as described above in Section 2. We analyzed their checklist to identify aspects that are covered or missing in AMSTAR 2 [42].

Molléri et al. [35] recently proposed a Catalog for Empirical Research in Software Engineering (CERSE) based on a systematic mapping study of 341 methodological papers that were identified using a combination of manual and snowballing search strategies. CERSE includes available guidelines, assessment instruments, and knowledge organization systems for empirical research in SE. To identify additional relevant articles that are not covered by Ampatzoglou et al. [7] in their tertiary

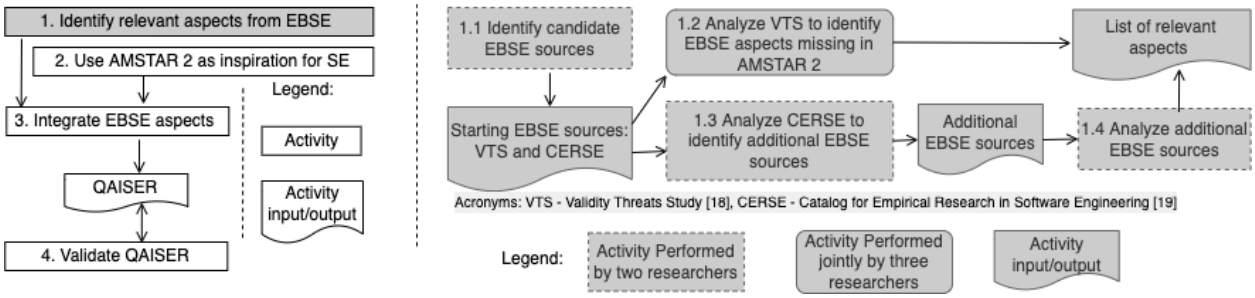

Figure 3: Step 1 - Identifying relevant aspects from EBSE literature.

study, we selected 74 articles from CERSE that are related to SLRs and mapping studies (SMSs). We obtained the source file containing the basic information (title of the paper, publication venue etc.) for these 74 articles from the first author of CERSE [35]. The first two authors independently reviewed these 74 articles to identify studies that propose or evaluate guidelines for conducting SLRs and SMSs in SE. Later, in a meeting, the first two authors developed a complete consensus on all 74 studies. The list of identified studies included, besides others, the latest version of the guidelines by Kitchenham et al. [31], the guidelines for mapping studies by Petersen et al. [37] and the guidelines for snowballing by Wohlin [48]. After including these three guidelines in our list of additional EBSE sources, we removed studies that were already covered in these guidelines [31, 37, 48].

Step 2: Using AMSTAR 2 as a source of inspiration for SE

The first two authors jointly analyzed AMSTAR 2 to identify items that are relevant for SE. As a validation, the third author independently reviewed the list of relevant and non-relevant items identified by the first two authors. Next, the first two authors adapted the response options for SE, for example, by replacing the medicine-specific options with the appropriate SE options. The adapted response options were also reviewed independently by the third author. After discussions, we achieved complete consensus between all three authors on all changes in items and response options.

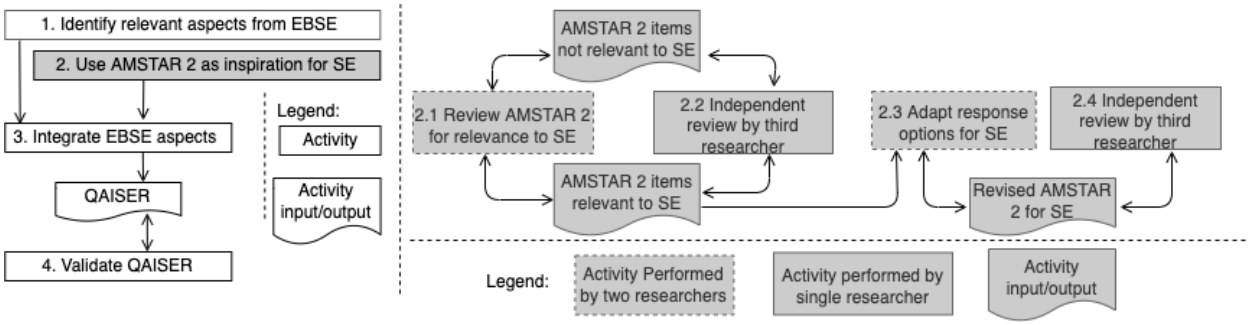

Figure 4: Step 2 - Using AMSTAR 2 as a source of inspiration for SE.

Step 3: Integrating EBSE aspects

Using the outputs of the previous steps and, in particular, the relevant EBSE literature identified in Step 1, the first two authors developed the first draft of QAISER. They also prepared a guidance document to support QAISER users in applying the instrument. The third author independently reviewed the instrument and the guidance document to validate its contents, i.e., to check that any

relevant aspect is not missed. The independent review helped improve the formulations and remove some inconsistencies in the instrument and the guidance document. However, it did not result in any significant change in the instrument.

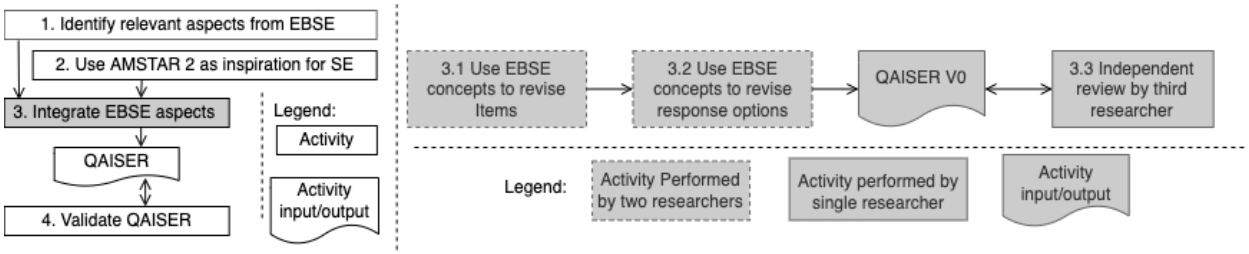

Figure 5: Step 3 - Integrating EBSE aspects.

Step 4: Validating QAISER

In this step, QAISER was reviewed by four experts in EBSE to validate the appropriateness of its items and reflect on its completeness (i.e., to identify if some aspects are missing) and understandability. The external experts are leading researchers in EBSE (see Table 2 for their profiles) and have been actively doing EBSE research since it was introduced in 2004 [32]. They have published several systematic secondary studies and made important methodological contributions to the EBSE discipline. In addition to QAISER and the guidance document, we prepared the following two documents to conduct the validation step (see Figure 6 for details about the validation step) systematically:

- A task description document: It described the steps that the external experts were asked to perform while reviewing QAISER. The task description document provided space where experts could enter their feedback on each QAISER item.
- A process description document: It briefly described the process we used to create QAISER.

Before the external validation, we performed a pilot validation with a senior colleague at our department who has experience of participating in multiple SLRs. The colleague reviewed all of the four documents mentioned above (i.e., task description, process description, QAISER, and the guidance document) and provided written feedback. We also conducted a follow-up interview (one hour, face-to-face) to discuss the feedback in detail and to ensure a shared understanding. We revised the task description and also the instrument based on the feedback collected during the pilot step. Most of the changes resulted in revised formulations. We shared the revised documents with our colleague and achieved consensus on the clarity of the task description and completeness and appropriateness of QAISER.

Next, we used the same approach with the external experts as we followed during the pilot. After obtaining the written feedback and performing the interviews (approximately one hour each and online) with all four external experts, we analyzed the comments to identify the changes that should be made in QAISER. Also, a revised version of QAISER and a summary of their feedback and our actions were sent to the external experts.

After the review of QAISER by external experts, we also applied it on a sample of six systematic literature studies (see Section 7 for more details, including how the six studies were selected). The application of QAISER resulted in some minor simplifications in the description of a few items. We also made our ratings of the six SLRs available online as additional support for using QAISER along with the guidance document and the QAISER instrument as a spreadsheet. Section 7 shares

#	Published systematic secondary studies	Methodological contributions to EBSE	h-index*
Expert 1	Several	Yes - introduced SLRs to SE, several major contributions to the SLR guidelines	80
Expert 2	Several	Yes - several contributions, including mapping study guidelines	41
Expert 3	Several	Yes - contributions to a few specific steps/phases within the SLR process	38
Expert 4	Several	Yes - contributions to a specific step/phase within the SLR process	28

* at the time of the review

Table 2: External experts' profiles

our reflections about using QAISER as a critical appraisal tool, and the links where the application related material is available online.

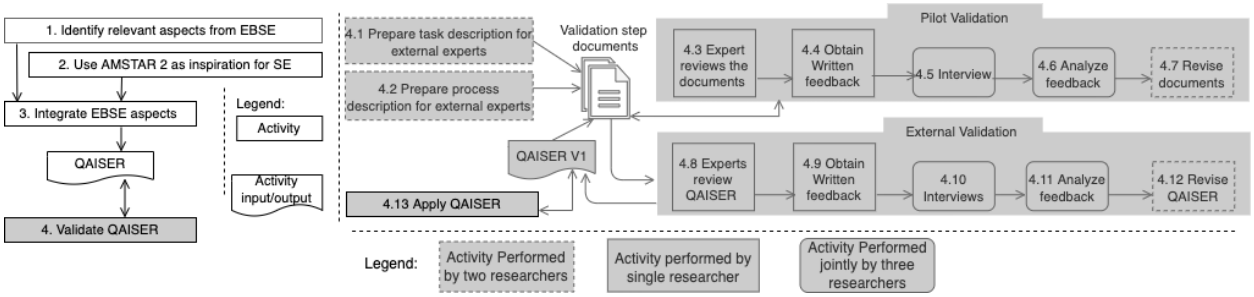

Figure 6: Step 4 - Validating QAISER.

4. Details of conducting the study

In this section, we present details of how we applied the process described in Section 3 and our justifications for the proposed changes to AMSTAR 2 while adapting it for SE.

In Section 4.1, we describe the development of the first version of QAISER (QAISER V0). This section is organized to explain the adaptations we made to AMSTAR 2 items, our justifications and the relevant EBSE sources used as the basis for the adaptations. For each AMSTAR 2 item, we start with the AMSTAR 2 item description, then provide our justifications for the proposed changes in the item, if any. QAISER V0 is not the final instrument. QAISER V0 was validated with the help of pilot and external validation steps. Section 4.2 presents the changes we made after the validation steps. In summary, Section 4.1 presents the output and the justifications of Steps 1-3 of the QAISER development process. Section 4.2 documents the changes to QAISER and their justifications made during Step 4 (Validating QAISER) of the QAISER development process.

4.1. Development of QAISER V0

In Step 1 of the process described in Section 3, we identified and selected four sources (see [7, 31, 37, 48]), in addition to DARE [11], from the EBSE literature to identify the relevant aspects for QAISER. Later, based on the suggestions of the external experts, we also included two more sources for identifying the relevant aspects for QAISER. The two additional sources related to a

framework for an automated-search strategy to improve the reliability of searches in SLRs [5], and a tertiary study describing lessons learned about reporting SLRs [8].

We now present the adaptation of AMSTAR 2 for SE based on the procedure detailed in Steps 2 and 3 of our method (see Section 3). Overall, at this stage in the process, we had two major changes in AMSTAR 2. The first change relates to the removal of existing items in AMSTAR 2. The removal includes excluding one item and replacing two items with a general item that is more appropriate for SE. The second change concerns the addition of an item.

In terms of removed items, three AMSTAR 2 items (Items 1, 11, and 12) were not included in QAISER as these were not deemed relevant to SLRs in SE. AMSTAR 2 Item 1 is about using PICO (Population, Intervention, Comparator group, Outcome) components in research questions and selection criteria. Items 11 and 12 are about meta-analysis, which is not commonly conducted in SE SLRs. We replaced these two items with a more general item about synthesis (see QAISER Item 11 in Appendix A). The new item checks if the included studies are synthesized or not. Synthesis of the included studies is one of the essential steps in an SLR [7, 8, 11, 31]. The details for these removed items are described later in the section.

The addition of one item is due to the following. Item 5 in AMSTAR 2 checks if the study selection is performed independently by at least two authors of the review. Item 6 checks the same aspect about the data extraction process. However, no item in AMSTAR 2 checks if the quality assessment is performed independently by at least two persons. We introduced an additional item to cover this aspect, i.e., to see if the quality assessment is performed independently by at least two authors of the review (see QAISER Item 10 in Appendix A).

We now describe in detail why and what changes were made to each item in AMSTAR 2. For each item, we initially state its AMSTAR 2 formulation and then explain the changes we proposed in it.

Item 1. ***"Did the research questions and inclusion criteria for the review include the components of PICO?"***
The previous guidelines [29] suggested the use of PICO to structure the research questions. However, the revised guidelines [31] excluded the suggestion for using this structured approach for research questions. The guideline authors noted that for SE reviews the use of this structured approach has not been found useful due to the lack of consistent and stable terminology, which makes it hard to derive relevant search keywords [31]. Due to such issues, PICO has not been widely used in SE (for details see: [16, 39]). The issue of reporting inclusion criteria is discussed in the changes to AMSTAR Item 3.
Changes: This item is not relevant for SE and was excluded from QAISER.

Item 2. ***"Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?"***
We identified no need to make any change at the item level. However, the following issues in the response options were noted:

a) The response options for 'Partial Yes' lack several aspects that are part of the protocol template included in the revised guidelines [31]. The missing aspects include description of the need for the review, data extraction process, synthesis process, threats to validity of the review, deviations from the protocol and the corresponding justifications for such deviations, and details of conducting the review.

b) Under 'Partial Yes', the authors are only required to state that they had a written protocol. The authors should also make the protocol publicly accessible and describe where and how can it be accessed [31].
c) One of the response options uses the term "risk of bias assessment". In SE, the more commonly used term is quality assessment.

Changes: Based on the analysis, the response options were modified as follows as an adaptation of them for SE:

a) The missing response options under 'Partial Yes' were added.
b) In the revised item, the authors are also required to make the protocol accessible and state how and where it can be accessed.
c) The risk of bias related response option was rephrased as quality assessment.

Item 3. ***"Did the review authors explain their selection of the study designs for inclusion in the review?"***
Most reviews in SE include different types of empirical studies. Thus, it is not relevant to ask for a justification for including all types of study designs. Furthermore, the study design is only one of the criteria for including or excluding studies from an SLR. Therefore, the item should address the larger aspect of the appropriateness of the inclusion and exclusion criteria. Reporting of the inclusion and exclusion criteria is also part of the DARE criteria [11] used by the Centre for Reviews and Dissemination at the University of York. Also, reporting of the inclusion and exclusion criteria and the relevant justifications is part of the guidelines [31] and other EBSE literature as well [7].
Changes: The item is reformulated as follows for SE: *Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?*
To get a 'Yes' score for the revised item, the review should have reported the inclusion and exclusion criteria and provided justifications for any restrictions used in the criteria.

Item 4. ***"Did the review authors use a comprehensive literature search strategy?"***
We identified the following issues in the response options:

a) The response options treat database search as the main search method while snowballing is partially addressed as an additional search method. In the revised guidelines for performing SLRs in SE [31], database and snowballing searches are included as alternate search strategies. Both strategies have been used in SE SLRs and have their own guidelines and best practices (for details see: [48, 4, 31]). In the current form of AMSTAR 2 Item 4 description, only the database search strategy could be assessed as comprehensive.
b) The response option related to the publication restrictions is more relevant to the inclusion and exclusion criteria.
c) Furthermore, two other response options are not used in SE SLR: The first one is about searching in the study registries, while the second one is about conducting the search within 24 months of completion of the review.

Changes: We introduced the following three changes:

a) Two groups of response options were created: first when a database search is used as the main search method and the second when a snowballing search is used as the main search method (See QAISER Item 4 in Appendix A) for details about the two groups of response options).
b) The response option related to the publication restrictions is moved to Item 3 (see Appendix A).

c) The two response options (searching in registries and search within last 24 months) were not included in QAISER. After the search is carried out, the review authors need to perform the remaining steps of the review process and write the review report. Finally, the review report is peer-reviewed before it can be published. The elapsed time in this entire process varies from case to case. To ensure the timeliness of the search, AMSTAR 2 includes this 24 months time limit, i.e., the search should have been conducted within the last 24 months. We removed the specific time limit of 24 months, which may not be appropriate in all cases. The appraisers are expected to judge the timeliness of the search (see Item 4 description in the guidance document in Appendix B) reported in the SLR under review.

Item 5. ***"Did the review authors perform study selection in duplicate?"***
We noted that:

a) The phrase "in duplicate" is not a commonly used term in SE and is therefore not self-explanatory. Furthermore, the item does not specify if the study selection is performed on the full text or on the titles and abstracts.
b) In the first response option, when all studies are reviewed independently by at least two authors of the review, the agreement level is not reported. Reporting of the agreement level would increase the transparency of the study selection process.
c) In the second response option, it is permitted that only a sample of the studies are independently reviewed by at least two authors of the review. The reliability of the study selection process is compromised if only a small sample of studies is reviewed by more than one author of the review. In particular, the excluded studies pose a threat to validity if a single person excludes them.

Changes: Three changes were introduced to address these observations:

a) The item was rephrased to clarify the focus on the independent study selection and that the initial study selection is based on titles and abstracts. The revised formulation is: *"Did the authors of the review* ***independently*** *perform study selection* ***based on titles and abstracts****?"*
b) At the end of the first response option, the following text is added to make it necessary to report the agreement level as well: *"... and reported the agreement level"*.
c) At the end of the second response option, the following text is added to make it compulsory to have the excluded studies reviewed by at least two authors: *"however, all excluded studies must be reviewed by at least two authors of the review"*.

Item 6. ***"Did the review authors perform data extraction in duplicate?"***
As in the previous item, the phrase "in duplicate" is not self-explanatory.
Changes: The item was rephrased in QAISER as follows: *"Did at least two authors of the review independently perform data extraction?"*

Item 7. ***"Did the review authors provide a list of excluded studies and justify the exclusions?"***
The item is about those studies that were excluded after reading the full text. The item does not indicate that it is about those studies that were read in full text, and not about those that were excluded based on the screening of the titles and abstracts.
Changes: The item was rephrased to indicate that it is about those studies that were read in full text. In the revised formulation, the following phrase is added at the end of the item text: *"...for*

the papers read in full text?"

Item 8. ***"Did the review authors describe the included studies in adequate detail?"***
The response options about intervention and outcomes may not be relevant to all SLRs in SE. In SE, not all SLRs would be about interventions and outcomes. The included studies in an SLR may not have investigated any interventions. Furthermore, not all studies in SE include human subjects. In such studies, the population may consist of other relevant items of interest such as artifacts, events, or some other aspects. We have clarified it further in the guidance document (Appendix B) when describing Item 8.
Changes: In the response options about interventions and outcomes, the phrase *"when applicable"* is added to explain that the review needs to describe only the relevant information about included studies.

Item 9. ***"Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?"***
We noted the following:

a) Other fields use RoB assessments that focus on methodological rigor and the impact of weaknesses on the reliability of results. Kitchenham et al. [33] in a recent study, clarify the difference between RoB and quality assessment as follows: *"The important difference between RoB and quality assessment for individual studies is that RoB is about identifying potential methodological flaws that can bias the outcome of primary studies, whereas quality is about whether the research was performed as well as possible."* In SE SLRs, the concept of quality assessment is more prevalent than RoB. Therefore, in QAISER, we have used the term quality assessment.
A variety of quality assessment instruments have been developed and used to assess the quality of the different types of empirical studies in SE [31]. The focus in SE so far has been on whether an SLR uses relevant quality assessment instruments. These instruments often cover both reporting quality and methodological rigor.
b) The current response options are not relevant to SE. Furthermore, the focus of the item is suggested to be changed to the quality assessment instrument. Therefore, the response options should also be revised accordingly to check the completeness and relevance of the questions in the quality assessment instrument.

Changes: We introduced the following changes:

a) We revised the item to emphasize whether or not the review authors have provided an explanation for their selection of the quality assessment instrument. The item is revised as follows: *"Did the review authors explain their selection of quality assessment instrument?"*
b) With regards to the response options under the revised item, for 'Yes', the review authors should have selected an appropriate quality assessment instrument for different types of studies included in the review. Furthermore, the instrument needs to have questions about study goals, research questions, appropriateness of the study design, data collection, and analysis methods. The instrument should also have question(s) about the study findings and the supporting evidence, and the extent to which the findings answer the research questions. We refer to the instrument in Appendix A for the specific response options for this item in QAISER.

Item 10. ***"Did the review authors report on the sources of funding for the studies included in the review?"***
This item focuses only on the sources of funding for individual studies. Funding is one of the

issues that could result in a conflict of interest. In some cases, the authors of the individual studies might have some other conflict of interest in favor of or against the topic or intervention they are investigating in their studies.
Changes: The item is revised to include any other conflict of interest besides funding sources. Conflict of interest is inserted in the item text as follows: *"Did the review authors report on the sources of funding* ***and any other conflict of interest*** *for the studies included in the review?"*

Item 11. ***"If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?"***
Meta-analysis studies are very rare in SE due to the lack of multiple empirical studies addressing the same research question. Therefore, this item is not relevant to the majority of the SE SLRs.
Changes: This item is removed from the adaptation of AMSTAR 2 for SE. We have instead included a more general item about synthesis (Item 11 in QAISER, see Appendix A).

Item 12. ***"If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?"***
As discussed with Item 11 above, meta-analysis is not common in SE SLRs. This item is removed from the adaptation of AMSTAR 2 for SE. However, it is important to note that considering the impact of the quality of individual studies while interpreting the results is still covered in the next item.

Item 13. ***"Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review?"***
We noted the following:

a) Instead of RoB, the SE community uses the notion of quality assessment more commonly.
b) The first response option deals with the inclusion of high-quality randomized controlled trials (RCTs). Since in SE, RCTs are not common, the focus should be on high-quality studies.
c) The second response option includes the requirement of discussing the impact of RoB on results. For SE, the focus has been on categorizing the analysis and interpretation of results based on study quality [31].

Changes: The following changes were introduced:

a) In line with Item 9 above, the RoB is replaced with quality of individual studies in the item description.
b) In the first response option, the phrase *"high quality RCTs"* is replaced with *"high quality studies"*.
c) The second response option is revised to focus on the categorization of the analysis and interpretation of results based on study quality.

Item 14. ***"Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity[1] observed in the results of the review?"***
We identified no need for adaptation to SE in this item.

[1] Heterogeneity occurs when the results are not consistent across studies. For example, different studies provide conflicting evidence for or against a SE intervention. It is important to investigate the causes of such inconsistent results before drawing any conclusions in such cases.

Item 15. ***"If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?"***

a) The item is limited to quantitative synthesis. In SE, qualitative synthesis is used more frequently in SLRs. Discussing publication bias and its impact on the results is essential, regardless of the type of synthesis performed, quantitative or qualitative. Publication bias may be that some authors have contributed with several studies within the area of the SLR, and it may affect the conclusions from the SLR. The latter becomes particularly critical if it is one or more of the researchers conducting the SLR.
b) The response option includes a requirement to carry out graphical or statistical tests as well. The main aspect to cover in this item should be to check if the authors of the review have discussed publication bias and discussed its potential impact on review results.

Changes: We introduced the following changes:

a) The item is made more general by removing the word quantitative while also adapting its formulation for SE.
b) The response option is also revised accordingly, i.e., removing the reference to the graphical or statistical tests. The revised response option aims to check if the publication bias and its impact on the results are discussed or not.

Item 16. ***"Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?"***
We identified no need for adaptation to SE in this item.

We call the resulting instrument that systematically adapts AMSTAR 2 for SE and supplements it with SE guidelines and evidence QAISER V0. This version was used in Step 4 (see Section 3 for details of the process) for further validation.

4.2. Changes in QAISER during the validation step

This section presents the changes made in QAISER V0 based on the feedback collected during the pilot and external validation steps.

Besides several editorial changes, the pilot validation resulted in the following two main changes in QAISER V0:

1) Addition of a new item on the need for undertaking the review (see QAISER Item 1 in Appendix A): In QAISER V0, establishing the need for undertaking a review was listed as one of the response options to score 'Partial Yes' under Item 1. During the discussions in the pilot validation, we agreed with the senior colleague to give more importance to establishing the need for the review step. The number of SLRs performed in SE is increasing every year. At times, there are multiple SLRs on the same topic. Thus, there is a need to establish if it is relevant to undertake a new SLR on a topic [34, 31, 37]. The authors of the review should justify the need for undertaking the review. To score 'Yes' on this new item in QAISER, the review should have 1) discussed the related existing reviews (if any) and established the need for another review by highlighting the gap, or 2) established the need to aggregate the evidence on a topic, if there exist no reviews on the topic.
2) Addition of a new response option under the synthesis related item in QAISER V0 (Item 11): Agreeing with the suggestion of the senior colleague, we added another response options under

Item 11 in QAISER to check how effectively the authors of the review have linked the answers and interpretations with the data extracted from the primary studies. The new response option is described as: *"Provided a clear trace linking the answers of review questions and interpretations to the data from the included primary studies"*.

The revised QAISER, after the pilot validation step, was shared with the external experts for further validation. The external experts provided several improvement suggestions. We provide a summary of the main suggestions related to items and response options in the following:

- Introduce an item about recommendations: SLRs are supposed to provide evidence-based input to practitioners and researchers to aid them in making informed decisions. QAISER did not have any item that specifically covered this aspect. The external experts suggested including an item that checks if the review provides appropriate recommendations and conclusions based on the review results. Agreeing with the external reviewer's suggestion, we added a new item about recommendations and conclusions in QAISER (see QAISER Item 14 in Appendix A).
- Remove the item about sources of funding (see AMSTAR 2 Item 10 described in Section 4.1): The item deals with the reporting of the sources of funding for the included studies. The external experts suggested to remove it as they did not find it relevant in SE context. We removed this item from QAISER.
- Reformulate Items 5 and 6: The external experts suggested to reformulate QAISER Items 5 and 6 to describe them at an appropriate level. In the current formulations of Items 5 and 6, the requirement to include two authors was included both in the descriptions of the items as well as in response options. In line with the suggestion, we reformulated both items as - *"Did the authors of the review include a reliable study selection (or data extraction) process?"*.
- Introduce 'Partial Yes' scale: Some items (Items 1, 5, 6, and 10) had a binary Yes/No scale. The external experts suggested introducing a third scale value of 'Partial Yes' to make them more flexible. We introduced a 'Partial Yes' option under these items and included the minimum acceptable requirements as response options (see QAISER Items 1, 5, 6, and 10 in Appendix A). AMSTAR 2 Items 5 (study selection) and 6 (data extraction) allow a 'Yes' rating even if multiple researchers were involved in reviewing only a sample of studies, with the rest being reviewed by only a single researcher. In the corresponding QAISER Items 5 and 6, such an approach (i.e., involvement of at least two researchers only on a sample of studies) results only in a 'Partial Yes' rating. QAISER Items 5 and 6 have more stringent requirements of involving at least two researchers to review the eligible or included studies for a Yes rating (see QAISER Items 5 and 6 in Appendix A).
- Quality focus: Assessing SLRs is not only about the presence or absence of an aspect; it is largely a subjective judgment concerning decisions and measures taken by the authors. To incorporate this suggestion, we introduced adjectives such as adequately, reliable, and appropriate in several items to assess SLRs' subjective nature better.
- Modifications to the protocol-related item (see AMSTAR Item 2 described in Section 4.1): The external experts suggested simplifying the response options for the 'Partial Yes' scale. We moved justification of any deviations from the protocol from 'Partial Yes' to the 'Yes' scale. Furthermore, threats to validity and details of conducting the review were removed from the 'Partial Yes' scale. We also removed a response option about heterogeneity from the 'Yes' scale. It was not deemed a necessary part of a protocol by the experts (see the revised description of QAISER Item 2 in Appendix A).
- Modifications to the heterogeneity-related item (see AMSTAR Item 14 described in Section 4.1): The external experts did not find this item to be essential for the systematic reviews in SE. The item is more relevant for meta-analysis studies, which are not common in SE. We replaced

the heterogeneity concept with the characteristics of the primary studies. Some differences in the results of the primary studies may be due to the variations in the studies' characteristics, e.g. if the participants in different studies are students or practitioners. Therefore, in the case when there are differences in the results of the primary studies, the authors of the review should perform an analysis to see if the differences are due to the variations in the primary studies' characteristics.

4.3. Concluding remarks

QAISER aims to support appraisers of SLRs in SE by raising important questions about the reliability and the relevance of an SLR. Furthermore, by providing evidence-based and accepted best practices in SE research (i.e., established expectations in the SE field of a high quality SLR), it supports the judgement of the conformance and the likely impact of non-conformance on the reliability and relevance of an SLR.

The quality aspects of concern and related criteria in QAISER are based on available evidence and recommendations in the SE literature. Therefore, the availability of evidence and the specificity of guidelines is also reflected in the criteria used in QAISER. Thus, the responses in the instrument range from specific/concrete actions to broader/general suggestions/guidelines. QAISER supports appraisers in making a judgement about the overall reliability and relevance of an SLR.

5. QAISER as an appraisal instrument

QAISER has three levels of judgement: item level, group level, and SLR level. It should be noted that AMSTAR 2 does not include these three levels. The levels are introduced to support the appraiser in moving towards an overall assessment of an SLR. However, the levels do not imply that QAISER aggregate the overall assessment to a final numeric score. The use of a single aggregate numeric score to compute and reflect on the quality of an SLR is not a recommended practice anymore [42] - instead a subjective assessment on an ordinal scale is preferred (e.g., high, medium, low, ad critically low ratings in AMSTAR 2 for the overall confidence in review results).

In this section, the three levels are presented in Section 5.1 (item level), Section 5.2 (group level) and Section 5.3 (SLR level) respectively.

Some items are more closely related to each other - e.g., Items 3, 4, 5, and 7 relate to the identification and selection of potentially relevant studies. Therefore, to allow appraisers to reflect on the strengths and weaknesses of the SLR in a group of related items in one place, we introduced the concept of group level assessment. After performing the item level assessment, appraisers perform the group level assessment, allowing them to assess the SLR on a group of related items. After this group level assessment, appraisers consolidate their assessment at the overall SLR level to judge if the SLR, as a whole, is reliable and relevant. At the SLR level, the assessments in the related groups support judging the relevance (two groups related the relevance: Groups 1 and 6) and reliability of the SLR (five groups related to the reliability: Groups 2, 3, 4, 5, and 7). Table 3 presents the items and the groups of QAISER, while the complete instrument and the guidance document are presented in Appendix A and B respectively (see supplemental material).

5.1. QAISER: item level assessment

The first level comprises 15 items formulated as questions. These questions are ordered to reflect the sequence of phases in the design, conduct, and reporting of a typical SLR. The criteria to meet the

Table 3: QAISER items and groups.

Group	Item description and the relevant sources/references
1. Motivation	Item 1: Did the authors of the review adequately justify the need for undertaking the review? [31, 37, 7]
2. Plan	Item 2: Did the authors of the review establish a protocol prior to the conduct of the review? [42, 7, 31]
3. Identification and selection	Item 3: Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions? [11, 31, 7] Item 4: Did the authors of the review use a comprehensive literature search strategy? [42, 31, 7, 11] Item 5: Did the authors of the review use a reliable study selection process? [42, 31, 37] Item 7: Did the authors of the review discuss and justify the exclusion of the potentially relevant studies that were read in full text? [42, 31]
4. Data collection and appraisal	Item 6: Did the authors of the review use a reliable data extraction process? [42, 31, 7, 37] Item 8: Did the authors of the review provide sufficient primary studies' characteristics to interpret the results? [42, 31, 7, 11] Item 9: Did the authors of the review use an appropriate instrument for assessing the quality of primary studies that were included in the review? [42, 31] Item 10: Did the authors of the review use a reliable quality assessment process? [31]
5. Synthesis	Item 11: Were the primary studies appropriately synthesized? [11, 31, 7, 8] Item 12: Did the authors of the review investigate the impact of the quality of individual studies on the results of the review? [42, 31, 7] Item 13: Did the authors of the review investigate the impact of primary studies' characteristics on the results of the review? [42, 31]
6. Recommendations and conclusions	Item 14: Did the authors of the review provide appropriate recommendations and conclusions from the review? [8]
7. Conflict of interest	Item 15: Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review? [42, 31]

questions on the item level are stated in the form of acceptable responses for each of the questions. All items are evaluated on a scale with two values (Yes/No) or three values (Yes/Partial Yes/No), i.e., an assessment of the extent to which an SLR under review fulfils the stated criteria.

Each item in QAISER is formulated with the objective that it is self-contained and self-explanatory. However, there is an accompanying guidance document (Appendix B in the supplemental material) with a more detailed description of the items and their responses. We recommend that before applying QAISER, the guidance document should be read, at least, before using QAISER for the first time.

5.2. QAISER: group level assessment

The external experts also provided a suggestion about clarifying the flow and sequence of the items in QAISER. To make the flow of the items more explicit and understandable, and to aggregate individual items into a logical cluster, we organized the 15 QAISER items into seven groups corresponding to the process and outcome of an SLR (see the first column in Table 3): (1) motivation to conduct a review, (2) plan and its validation, (3) identification and selection, (4) data collection and quality appraisal, (5) synthesis, (6) recommendations and conclusions, and (7) conflict of interest.

At the group level, the assessment results on the item level are used as indicators for major and minor weaknesses based on their impact on the reliability and relevance of an SLR, see Table 4. Having completed the assessment of individual QAISER items, an appraiser should reflect on the impact of the weaknesses on the reliability and relevance of the SLR at the group level. Groups

1, 2, 6, and 7 consist of single items only, and are therefore relatively simple to reflect upon. A "No" rating on the corresponding items of these four groups indicates a major weakness at the group level. Groups 3, 4, and 5 consist of multiple items and are more complex to reflect upon. The appraisers should make an overall assessment after considering the ratings of all items in the groups. As a rule of thumb, we recommend that all items receiving a "No" should be considered as hinting at a major weakness in the group being assessed.

Table 4: QAISER: group level assessment.

Group	**Item ranking (Yes/ Partial Yes /No)**	**Impact**	**Comments**
1. Motivation	Item 1 (need):		
2. Plan	Item 2 (protocol):		
3. Identification and selection	Item 3 (selection criteria): Item 4 (search): Item 5 (selection process): Item 7 (excluded studies):		
4. Data collection and appraisal	Item 6 (data extraction): Item 8 (study characteristics): Item 9 (quality criteria): Item 10 (quality assessment process):		
5. Synthesis	Item 11 (synthesis): Item 12 (considered study quality) Item 13 (considered study characteristics)		
6. Recommendations and conclusions	Item 14 (recommendation):		
7. Conflict of interest	Item 15 (their own):		

5.3. QAISER: SLR level assessment

By progressively building on the first two levels, an appraiser judges the overall reliability and relevance of an SLR at the SLR level. Thus, considering the impact of weaknesses in related groups, i.e., **relevance** (mainly two groups: motivation, and recommendations and conclusions) and **reliability** (mainly the following five groups: plan, identification and selection, data collection and appraisal, synthesis, and conflict of interest).

AMSTAR 2 suggests that the appraisers should pre-specify which items are more critical for an SLR under review. They have also suggested an advisory list of critical items. For QAISER, we recommend a similar approach. The appraisers should pre-specify which groups of items are more/less critical for the specific review being assessed.

In the following text, we provide two examples of using the group-level assessment for assessing the reliability and relevance of an SLR.

Interpretation criteria, example 1:

1) Reliability of an SLR: The reliability of an SLR is assessed by rating the overall confidence in the results of an SLR as: high, moderate, low or critically low. Apart from group 1, all other groups are relevant while considering the confidence in the results of an SLR. As a rule of thumb, we recommend that the confidence in the SLRs with major weaknesses in groups 3, 4, and 5 should be considered "critically low".

Table 5 provides guidance for interpreting weaknesses observed at the group level to select a confidence rating at the SLR level.

2) Relevance of an SLR: The relevance of an SLR is also rated as high, moderate, low or critically low. Groups 1 and 6 are considered when making a judgement about the relevance of an SLR. As a rule of thumb, we recommend that the relevance of an SLR be judged to be "critically low" if there are major weaknesses in both groups 1 and 6. Table 6 provides guidance in selecting a relevance rating based on the weaknesses in groups 1 and 6.

Interpretation criteria, example 2: A more stringent criteria for assessing reliability (instead of Table 5) could be stated as follows:

– Critically low – major weakness in at least one of the groups 3, 4, 5, or 7.
– Low – several minor weaknesses in groups 3 and 4, and 5
– Moderate – no major weakness in groups 3, 4, 5, and 7, but a major weakness in groups 2 or 6
– High – only minor weaknesses in at most two of the groups 3, 4, 5 and only a few minor weaknesses in groups 2, 6, and 7

The above criteria acknowledge that a major limitation in any one of the groups 3, 4, 5, or 7 i.e., search and selection (group 3), data collection and appraisal (group 4), synthesis (group 5), or conflict of interest (group 7) cannot be compensated for by excellence in other groups. For example, a thorough synthesis does not compensate for limitations in the search.

Table 5: Judging the reliability of the results of an SLR.

Reliability	**Suggestions for interpretation based on group-level assessment in Table 4**
"critically low"	– major weaknesses in groups 3, 4, and 5 –
"low"	– major weaknesses in at most two of the groups 3, 4, 5 along with major weaknesses in groups 2 and 6 –
"moderate"	– no major weakness in groups 3, 4, 5, and 7, but major weaknesses in groups 2 or 6 –
"high"	– only minor weaknesses in at most two of the groups 3, 4, 5 and only a few minor weaknesses in groups 2, 6, and 7 –

Table 6: Judging the relevance of an SLR.

Relevance	**Suggestions for interpretation based on group-level assessment in Table 4**
"critically low"	– major weaknesses in group 1 and 6 –
"low"	– major weakness in either group 1 or 6 –
"moderate"	– minor weaknesses in both groups 1 and 6 –
"high"	– only a minor weakness in group 6 –

6. Reliability of QAISER

In this section, we highlight three aspects that contribute to the reliability of QAISER as a potentially effective instrument for assessing the quality of SLRs in SE.

1) The relevance of AMSTAR and AMSTAR 2 validations: The original AMSTAR [41] consisted of 11 appraisal items. Based on community feedback, AMSTAR 2 was proposed consisting of 16 items with an increased focus on the risk of bias evaluation and the possibility to assess SLRs that may have non-randomized studies. Both AMSTAR and AMSTAR 2 have been used and

validated extensively (for details see:[20, 38, 40]). These validation efforts provide credibility to QAISER as well, as most of its items (12 out of 15) are adapted from AMSTAR 2.

2) Comparison with DARE: DARE [11] is the most frequently used criteria to assess the quality of SLRs. Several essential aspects related to the quality of SLRs are not covered in DARE, e.g., justifying the need to conduct a review, establishing a protocol prior to performing the review, study selection process, data extraction process, and quality assessment process. Furthermore, three of the DARE criteria (including the important criterion about synthesis) are limited to checking the presence/absence of different aspects, rather than their appropriateness, e.g., *if the inclusion and exclusion criteria are reported or not?* QAISER not only covers aspects that are missing in DARE, but it also focuses on quality aspects of different criteria—for example, checking the appropriateness of inclusion and exclusion criteria rather than only focusing on the mere reporting of such criteria in the review report.
3) External validation. We followed a systematic process (see Section 3 for details) to adapt AMSTAR 2 for SE and to introduce additional aspects based on the recommendations in the systematically identified EBSE literature. Four leading experts then reviewed the proposed instrument to check the appropriateness, completeness, and understanding of its items (refer to Section 3 for details about the validation step). The experts recommended some changes, which we incorporated in the revised version of QAISER. The experts did not suggest any further changes in the revised version. Although QAISER is developed using a systematic process, the proposed changes while adapting AMSTAR 2 to SE still need to be empirically validated by independent researchers.

7. Application of QAISER

In this section, we describe the support available for applying QAISER and share our reflections on applying QAISER to assess a sample of six systematic literature studies. The selected studies include five systematic literature reviews and one mapping study [52] published in SE. None of the five SLRs used meta-analysis, and hence the adaptations made in QAISER for assessing more qualitative SLRs were tested. For example, [45] aimed at meta-analysis, but had to settle for vote counting, and [18] used three synthesis approaches: narrative synthesis, vote counting and reciprocal translation. Thus, the analysis includes two approaches being qualitative of nature. The other three systematic literature studies, [24], [17] and [6] also used more qualitative approaches to synthesize their findings, including, an adaptation of comparative analysis, and in the other two SLRs, categorisation was conducted by the authors. Thus, the five SLRs primarily use qualitative approaches to synthesize the findings.

7.1. Support for Applying QAISER

In line with AMSTAR 2, we also developed a guidance document (see Appendix B) for supporting appraisers in applying QAISER. The guidance document describes the following aspects for each QAISER item:

1) What is the item about? We provide a brief description of the item.
2) How to assess the item? We explain what an appraiser needs to check for assigning 'Partial yes', 'Yes', and 'No' ratings.
3) Where to find the relevant information to assess the item? We provide hints concerning which sections of the review papers are most likely to have the information needed to assess the item.

To further support the application of QAISER, we developed a spreadsheet to operationalize the QAISER instrument.

7.2. Reflections on Applying QAISER

We applied QAISER on a sample of six systematic literature studies (SLS) [18, 24, 45, 52, 17, 6] - five SLRs and one SMS. The six systematic studies were selected from four tertiary studies. The four tertiary studies were selected based on being published recently (2018-2020) and using DARE with five criteria. The purpose of applying QAISER on a sample of SLSs was to validate its usefulness - and also to make available the ratings of the selected SLSs as additional support for applying QAISER. The six SLSs were selected as follows:

- The second author selected three SLRs from the tertiary study by Kitchenham et al. [8] that were previously assessed using DARE with the highest ranking. The purpose of selecting the high-ranking SLRs on DARE criteria was to illustrate the usefulness of QAISER in supporting appraisers in performing a more fine-grained and thorough critical appraisal compared to DARE.
- The third author selected three SLSs, one from each of the other three tertiary studies meeting the criteria concerning publication year and using DARE with five criteria. One of the tertiary studies is also generic in a similar way as the tertiary study above [26]. However, it covers mapping studies. It was decided to include this tertiary study to evaluate DARE vs QAISER on a mapping study too. The other two tertiary studies cover different topics within SE [15, 10]. For each of these three tertiary studies, one systematic study was randomly selected. The DARE scores for the three selected SLSs were 2, 2.5 and 3 respectively. Thus, the objective was to also apply QAISER on SLSs with lower DARE scores than the highest score (5), i.e., to contrast with the three SLRs selected by the second author.

We assessed the six selected SLSs with QAISER in two pairs of raters: 1) the first and the second author for the three SLSs selected by the third author, and 2) the first and the third authors for the three SLRs selected by the second author. Each QAISER item was used six times - on six different SLSs. The assessment resulted in the application of 15 QAISER items on six selected SLSs by the two pairs of raters - i.e., in total 90 assessment decisions for six SLSs. For a majority of the items, there were either no or minor differences (11 minor differences in total of the type Yes/Partial-Yes or Partial-Yes/No). There were also 12 differences of the type Yes/No in total - for no item we had more than two major differences of Yes/No type. Looking at both types of differences - minor and major - we noticed that for Item 1 (motivation), Item 2 (protocol) and 14 (Recommendations), we had relatively more differences as compared to the other items. Nevertheless, all authors found the option to add comments extremely helpful in reflecting on our ratings and corresponding justifications. Besides sharing raters' perspective, the comment fields also provide an opportunity to reflect on those aspects that are not captured in the QAISER instrument.

We also noted that the possibility to build on the item level assessment at the group and overall SLR level helped us in arriving at an overall rating for the SLR in a subjective, but informed and structured way. Both pairs of raters arrived at similar ratings at the SLR level, i.e., there were no major disagreements. For the relevance judgement, three SLSs received exactly the same rating, while the remaining three had minor differences as follows: high/moderate ratings (two SLSs) and low/moderate ratings (one SLS). Likewise for the confidence judgment as well, three SLSs received exactly the same ratings, while the remaining three have minor differences as follows: low/critically low ratings (two SLSs), low/moderate ratings (one SLS).

We observed that the time spent on QAISER application is relatively more when it is used for the first time. But for the second and third application, all authors noted that we were able to

complete the assessment relatively quickly. In case of the tertiary studies wherein the quality of several SLRs needs to be assessed, the authors would have to spend some time in the start - on studying the QAISER instrument and the guidance document before the first time use - but can then continue to use it on the remaining SLRs more efficiently. Using QAISER instead of DARE would relatively be more time-consuming. However, QAISER provides support in performing a more fine-grained and thorough critical appraisal compared to DARE. Even for those aspects that are already covered in DARE (e.g., inclusion and exclusion criteria and search strategy), we were able to perform a more fine-grained appraisal. The support of QAISER worked equally well in assessing the systematic mapping study with one expected difference related to the time spent on the assessment. The assessment of the mapping study [6] with QAISER was relatively less time-consuming as we did not have to apply the QAISER items related to the quality assessment and synthesis (i.e., Items 9 to 12).

The guidance document, QAISER instrument, and the spreadsheet corresponding to the six example applications are all available online [2]. Researchers could look at the six examples as additional support complementing the guidance document.

Based on our experiences of applying QAISER on the selected systematic studies, we decided to make the following minor simplifications in the QAISER instrument:

- Item 2 (Protocol): In the previous version, there was a requirement that the protocol should be publicly available. For the six SLSs we assessed, we noted that most authors included the necessary parts of the protocol in the review report. We made a small change in the instrument that allows authors to either include the protocol in the review report or make it available online.
- Item 15 (Conflict of interest): In the previous version, this item explicitly covered reporting of the funding sources and any other potential sources of conflict of interest. The guidance document covered the possible impact of the review authors' own prior research on the review results while explaining "any other potential sources of conflict of interest". We noted that financial conflicts of interest were irrelevant for most studies during the assessment of the six selected SLSs. However, the potential impact of review authors' own prior research on the review results became apparent. Therefore, we made a minor change in this item by lifting this consideration from the guidance document and making it part of it. The item scale has been slightly adjusted to reflect the renewed focus.

Some QAISER items and response options were also reformulated during the peer review process based on the recommendations of the reviewers. In addition, the review process also resulted in a few minor changes in the guidance document. The changes in the descriptions of the QAISER items during the review process include the following:

- To link the inclusion and exclusion criteria of the review with its research questions, Item 3 is revised as: "*Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?*"
- To further clarify the expectation from the authors of the review to analyze the impact of the quality of the included studies on the review results, Item 12 is slightly modified as: "*Did the authors of the review investigate the impact of the quality of individual studies on the results of the review?*"
- To further clarify the expectation from the authors of the review to analyze the impact of the characteristics of the included primary studies on review results, Item 13 is also reformulated:

[2] https://drive.google.com/drive/folders/1p7OUEfqQTF4dY3e_OX_OHiyi_tC4E_cU?usp=sharing

"*Did the authors of the review investigate the impact of primary studies' characteristics on the results of the review?*"
– Item 15 is also simplified as: "*Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?*"

The final version of the QAISER is presented in Appendix A.

8. Threats to validity

In this research, we aimed to propose an instrument for appraising systematic literature reviews in SE. In the design and conduct of this research, the following two objectives guided us:

1) To develop an instrument that is comprehensive, practical and appropriate for SE. Thus, the instrument shall cover all essential elements concerning the quality of an SLR, assist the appraiser when judging the quality of an SLR, and take into account the SE body of knowledge.
2) To reduce the researcher's bias in the development of the instrument.

The two main threats to validity identified concerned researcher bias and applicability of QAISER. The researchers come from the same affiliation, which creates a risk of having a coherent view on research. When creating an instrument for use by the research community, there is a risk that the instrument is hard to understand, and hence limiting its applicability.

To achieve the two objectives above and mitigate the threats to validity, we undertook the following actions:

– **Benefits and limitations of using AMSTAR 2 as a foundation for QAISER:** We used AMSTAR 2 as the starting point for our work as it is a well used and validated tool [5]. Albeit, AMSTAR 2 is primarily developed for quantitative studies as shown through Item 11 and 12, which focus on meta-analysis. However, as qualitative studies and mapping studies are more common in SE than quantitative studies using meta-analysis, we have adapted such items from AMSTAR 2 in QAISER to match the software engineering context. The adaptation was assessed through the external experts and applying QAISER to five SLRs and one mapping study from SE. None of the six studies used meta-analysis. As a consequence, items inherited from AMSTAR 2 have been validated to a larger extent than the newly introduced items in QAISER (Items 1, 10, and 14) and the adaptations to SE.
– **A systematic and rigorous process:** As described in Section 3, we followed a systematic approach for the development of QAISER. All data collection, analysis and interpretation involved at least two researchers. A third researcher independently reviewed the outcomes from several individual phases in the study. We maintain traceability for the adaptations in the existing tool by documenting the reasons, sources consulted and the changes.
– **Validation with leading experts in the field:** QAISER was reviewed by only four external experts, which is a limitation of our study. The number of experts is limited to four, but they include some of the main contributors of the methodological guidelines for designing, conducting and reporting secondary studies in SE. They have also authored several SLRs and have conducted several studies reporting critical evaluations of existing SLRs in SE.
– **Capturing agreed design, process and reporting guidelines in QAISER:** QAISER aims to support appraisers in assessing the quality of completed SLRs. As a critical assessment instrument, it ought to reflect the current best practices and established guidelines regarding the systematic review design, process, and reporting in software engineering. For this purpose, we used a systematic approach to identify and select a representative and current list of sources

to capture the current best practices and evidence for inclusion in QAISER (see Step 1 in Section 3 for details of our approach and the sources considered when developing QAISER). Thus, we minimized the risk of overlooking relevant literature or proposing an instrument in conflict with the current best practices by undertaking the following actions: (1) a thorough search utilizing two recent and comprehensive reviews of methodological guidelines [7, 35], (2) the inclusion of key guidelines [31, 37], and (3) consulting experts who are the main contributors to the methodological guidelines in software engineering.

With these actions, we have tried to mitigate the major threats to the validity of QAISER. However, the instrument needs to be used by other researchers beyond the authors of QAISER. Similar use and validation of AMSTAR identified several necessary improvements and resulted in AMSTAR 2. Such an evaluation of the instrument is essential to form a basis for further improvement of QAISER related to aspects like usability and reliability. QAISER introduced the concept of item, group, and SLR level assessments. This multi-level assessment was introduced to support appraisers in arriving at an overall assessment at the SLR level. Apart from the six SLSs assessed by the authors, this multi-level assessment mechanism has not been validated by external researchers. We plan to conduct external validation as part of our future research. We have made QAISER publicly available on online research forums [3]. In this paper, the idea is to publish QAISER with the current validation - wherein QAISER has been reviewed by a few experts (during pilot testing and external review) and applied by the authors to a sample of SLRs - and then continue to update it based on the community feedback and evidence from the future external validation studies.

By making QAISER and guidance for its usage publicly available, we hope that we and others in the field will address this need in the future.

9. Discussion

Given the lack of an appraisal instrument for assessing the quality of SLRs in SE, we developed QAISER. As presented in the introduction (see Section 1), researchers in SE have used the criteria in DARE [11] for assessing the quality of SLRs, although it comes with limitations [4]. Furthermore, to simply use an appraisal instrument, such as AMSTAR 2, from another discipline also comes with issues as illustrated in the development of QAISER. There was a need to adapt AMSTAR 2 to SE, and hence AMSTAR 2 is not an option by itself. The differences between disciplines need to be captured in the appraisal instrument.

QAISER takes its starting point from a well-established appraisal instrument from another field, i.e., AMSTAR 2 from the field of evidence-based healthcare. Furthermore, QAISER incorporates best practices from the conduct of SLRs in SE. Thus, QAISER is well-grounded in the literature and contributes to taking the quality assessment of SLRs in SE one step forward.

The objective of QAISER is to support appraisers of SLRs in SE. The expertise of the individual appraisers is crucial, and it cannot be replaced with an appraisal instrument such as QAISER.

The DARE assessment criteria are used to aggregate the ratings on the five questions as a final numeric quality score (maximum value is five - corresponding to five DARE questions). QAISER, like AMSTAR 2, does not assign a final numeric score for quantifying the quality of an SLR. On the contrary, QAISER is intended to support appraisers by covering the most salient quality aspects of an SLR. Thus, QAISER will help identify major and minor weaknesses in the design, execution, or reporting of an SLR that compromise the SLR results' relevance and our confidence in them.

[3] QAISER is publicly available on ResearchGate https://www.researchgate.net/publication/354765980_A_Quality_Assessment_Instrument_for_Systematic_Literature_Reviews_in_Software_Engineering and on arXiv https://arxiv.org/abs/2109.10134

Although the main objective is to support appraisers in assessing the quality of completed SLRs, we believe that authors of SLRs may also use QAISER to help them with improving the quality of their SLR before submitting the research for assessment. In the best of worlds, each submitted SLR is of high quality at the submission stage. It should be noted that the quality of an SLR is highly influenced by the quality of the primary studies included. The need to assess the quality of primary studies is highlighted by, for example, Dybå and Dingsøyr [19], and Yang et al. [51]. With the same objective, Wohlin highlights the need to write for synthesis when publishing primary studies [49].

We recommend all users of QAISER look at not only the appraisal instrument itself but also the accompanying guidance document. The latter is particularly important when using the instrument for the first couple of times. We have also made available online a spreadsheet operationalizing QAISER and six example assessments to further support appraisers in using QAISER.

The items and their response options in QAISER are intended to help highlight areas with weaknesses (or room for improvement). Given that assessment is prone to bias, we have deliberately chosen to have two or three levels for assessing each item. More levels may increase the risk for appraiser bias, although it may also benefit since the scale becomes more fine-grained. However, since QAISER is geared towards supporting appraisers of SLRs, we leave it to each appraiser to tune the feedback in writing using the comments option provided with each item, rather than having a more fine-grained scale.

When using QAISER for a mapping study, some items or questions may be less applicable than for an SLR, for example, the item concerning synthesis. We did consider adding an option of "not applicable" for mapping studies. We have chosen not to make the appraisal instrument more complex by adding the "not applicable" option. Thus, we leave it to each appraiser to decide if something is not applicable for a mapping study. Our preference is to leave freedom to the appraiser, given that SLRs and mapping studies may come in different shapes and colors. Assessing SLRs and mapping studies is a subjective endeavour, and the objective of any appraisal instrument should be to support the expert appraiser.

10. Conclusion and Future Work

QAISER, as an appraisal instrument for SLRs in SE, is built on a well-established appraisal instrument from another discipline (AMSTAR 2), and a set of guidelines, checklists, and experiences from SE. Furthermore, four external experts on SLRs in SE have reviewed an earlier version of QAISER, and QAISER has been revised based on their feedback. QAISER has also been used to assess the quality of six selected SLRs to demonstrate its applicability. Thus, QAISER is well-founded, and hence it is ready for further validation through usage.

QAISER includes 15 items and several response options for each item to assess for appraisers to arrive at an assessment for each item. QAISER provides support to consolidate the items on a group level, which is not done in AMSTAR 2. In QAISER, the items are consolidated into seven groups to support the appraiser to get a good overview of the strengths and potential weaknesses of an SLR. Moreover, QAISER has support for consolidating from the group level to the SLR level. The assessment of each group is systematically used to form an opinion about the overall quality of an SLR both in terms of reliability and relevance. AMSTAR 2 only provides an overall assessment of the confidence in the results. Given the importance of both reliability and relevance of the results for SE, we have provided support for both aspects.

In the future, we plan to evaluate the reliability and usability of QAISER by asking independent researchers to use it to assess the quality of selected SLRs. Based on such feedback, we plan to enhance QAISER further to support the SE community in assessing SLRs.

Acknowledgements

We would like to express our sincere thanks to the external experts: Prof. Daniela S. Cruzes, Prof. Barbara Kitchenham, Prof. Stephen G. MacDonell and Prof. Kai Petersen for their constructive feedback on QAISER.

We would also like to thank Prof. Jürgen Börstler for his kind participation in the pilot of the study. His detailed feedback helped us to improve the planning and execution of the evaluations with external experts. We also extend our gratitude to Dr. Jefferson S. Molléri for providing the listing of articles from the work with CERSE. Lastly, we are also thankful to the anonymous reviewers for their value feedback, which has helped us to further improve the Manuscript.

This work has been supported by ELLIIT, a Strategic Area within IT and Mobile Communications, funded by the Swedish Government. The work has also been supported by research grants for the VITS project (reference number 20180127) and the OSIR project (reference number 20190081) from the Knowledge Foundation in Sweden.

References

[1] N. B. Ali, E. Engström, M. Taromirad, M. R. Mousavi, N. M. Minhas, D. Helgesson, S. Kunze, and M. Varshosaz. On the search for industry-relevant regression testing research. *Empirical Software Engineering*, 24(4):2020–2055, 2019.

[2] N. B. Ali and K. Petersen. Evaluating strategies for study selection in systematic literature studies. In *Proceedings of ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM*, pages 45:1–45:4. ACM-IEEE, 2014.

[3] N. B. Ali, K. Petersen, and C. Wohlin. A systematic literature review on the industrial use of software process simulation. *J. Syst. Softw.*, 97:65–85, 2014.

[4] N. B. Ali and M. Usman. Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy. *Inf. Softw. Technol.*, 99:133–147, 2018.

[5] N. B. Ali and M. Usman. A critical appraisal tool for systematic literature reviews in software engineering. *Inf. Softw. Technol.*, 112:48 – 50, 2019.

[6] D. Ameller, X. Burgués, O. Collell, D. Costal, X. Franch, and M. P. Papazoglou. Development of service-oriented architectures using model-driven development: A mapping study. *Information and Software Technology*, 62:42–66, 2015.

[7] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Inf. Softw. Technol.*, 106:201–230, 2019.

[8] D. Budgen, P. Brereton, S. Drummond, and N. Williams. Reporting systematic reviews: Some lessons from a tertiary study. *Inf. Softw. Technol.*, 95:62 – 74, 2018.

[9] B. U. Burda, H. K. Holmer, and S. L. Norris. Limitations of a measurement tool to assess systematic reviews (AMSTAR) and suggestions for improvement. *Systematic reviews*, 5(1):58, 2016.

[10] H. Cadavid, V. Andrikopoulos, and P. Avgeriou. Architecting systems of systems: A tertiary study. *Information and Software Technology*, 118:106202, 2020.

[11] Centre for Reviews and Dissemination, University of York. Database of abstracts of reviews of effects (DARE). https://www.crd.york.ac.uk/CRDWeb/AboutPage.asp, 2019. 29 Nov, 2019.

[12] N. Condori-Fernandez, R. J. Wieringa, M. Daneva, B. Mutschler, and O. Pastor. An experimental evaluation of a unified checklist for designing and reporting empirical research in software engineering. Technical report, Centre for Telematics and Information Technology (CTIT), The Netherlands, 2012.

[13] D. Costal, C. Farré, X. Franch, and C. Quer. How tertiary studies perform quality assessment of secondary studies in software engineering. In *Proceedings of the XXIV Iberoamerican Conference on Software Engineering*, page 1. Curran Associates Inc., 2021.

[14] D. S. Cruzes and T. Dybå. Research synthesis in software engineering: A tertiary study. *Inf. Softw. Technol.*, 53(5):440 – 455, 2011.

[15] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli. Usability in agile software development: A tertiary

study. *Computer Standards & Interfaces*, 64:61–77, 2019.

[16] F. Q. Da Silva, A. L. Santos, S. C. Soares, A. C. C. França, and C. V. Monteiro. A critical appraisal of systematic reviews in software engineering from the perspective of the research questions asked in the reviews. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 1–4. IEEE – ACM, 2010.

[17] M. Daneva and B. Lazarov. Requirements for smart cities: Results from a systematic review of literature. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–6. IEEE, 2018.

[18] O. Dieste and N. Juristo. Systematic review and aggregation of empirical studies on elicitation techniques. *IEEE Transactions on Software Engineering*, 37(2):283–304, 2010.

[19] T. Dybå and T. Dingsøyr. Strength of evidence in systematic reviews in software engineering. In H. D. Rombach, S. G. Elbaum, and J. Münch, editors, *Proceedings of the Second International Symposium on Empirical Software Engineering and Measurement, ESEM*, pages 178–187. ACM, 2008.

[20] A. Gates, M. Gates, G. Duarte, M. Cary, M. Becker, B. Prediger, B. Vandermeer, R. M. Fernandes, D. Pieper, and L. Hartling. Evaluation of the reliability, usability, and applicability of AMSTAR, AMSTAR 2, and ROBIS: protocol for a descriptive analytic study. *Systematic reviews*, 7(1):85, 2018.

[21] T. Greenhalgh. How to read a paper: Papers that summarise other papers (systematic reviews and meta-analyses). *BMJ*, 315(7109):672–675, 1997.

[22] J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, and V. Welch. *Cochrane handbook for systematic reviews of interventions*. Wiley, January 2019.

[23] M. Höst and P. Runeson. Checklists for software engineering case study research. In *Proceedings of the first International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 479–481. IEEE, 2007.

[24] A. Idri, F. azzahra Amazal, and A. Abran. Analogy-based software development effort estimation: A systematic mapping and review. *Information and Software Technology*, 58:206–230, 2015.

[25] K. S. Khan, G. Ter Riet, J. Glanville, A. J. Sowden, J. Kleijnen, et al. Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews. Technical Report 4 (2n, University of York, UK, 2001.

[26] M. U. Khan, S. Sherin, M. Z. Iqbal, and R. Zahid. Landscaping systematic mapping studies in software engineering: A tertiary study. *Journal of Systems and Software*, 149:396–436, 2019.

[27] B. Kitchenham. Procedures for Performing Systematic Reviews. Technical report, Keele University, Keele, UK, 2004.

[28] B. Kitchenham and P. Brereton. A systematic review of systematic review process research in software engineering. *Inf. Softw. Technol.*, 55(12):2049 – 2075, 2013.

[29] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. Technical report, School of Computer Science and Mathematics, Keele University, Keele, UK, Keele, UK, 2007.

[30] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering – A tertiary study. *Inf. Softw. Technol.*, 52(8):792–805, Aug. 2010.

[31] B. A. Kitchenham, D. Budgen, and P. Brereton. *Evidence-based software engineering and systematic reviews*, volume 4. CRC press, 2015.

[32] B. A. Kitchenham, T. Dyba, and M. Jorgensen. Evidence-based software engineering. In *Proceedings. 26th International Conference on Software Engineering*, pages 273–281. IEEE, 2004.

[33] B. A. Kitchenham, L. Madeyski, and D. Budgen. SEGRESS: Software Engineering Guidelines for REporting Secondary Studies. *IEEE Transactions on Software Engineering*, pages 1–1, 2022.

[34] E. Mendes, C. Wohlin, K. R. Felizardo, and M. Kalinowski. When to update systematic literature reviews in software engineering. *J. Syst. Softw.*, 167:110607, 2020.

[35] J. S. Molléri, K. Petersen, and E. Mendes. CERSE-Catalog for empirical research in software engineering: A systematic mapping study. *Inf. Softw. Technol.*, 105:117–149, 2019.

[36] I. Nurdiani, J. Börstler, and S. A. Fricker. The impacts of agile and lean practices on project constraints: A tertiary study. *J. Syst. Softw.*, 119:162–183, 2016.

[37] K. Petersen, S. Vakkalanka, and L. Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.*, 64:1 – 18, 2015.

[38] D. Pieper, R. B. Buechter, L. Li, B. Prediger, and M. Eikermann. Systematic review found AMSTAR, but not r (evised)-AMSTAR, to have good measurement properties. *J. Clin. Epidemiol.*, 68(5):574–583, 2015.

[39] M. Riaz, M. Sulayman, N. Salleh, and E. Mendes. Experiences conducting systematic reviews from novices' perspective. In *14th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 1–10. BCS Learning & Development Ltd., Swindon United Kingdom, 2010.

[40] B. J. Shea, L. M. Bouter, J. Peterson, M. Boers, N. Andersson, Z. Ortiz, T. Ramsay, A. Bai, V. K. Shukla, and J. M. Grimshaw. External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLoS One*, 2(12):e1350, 2007.

[41] B. J. Shea, J. M. Grimshaw, G. A. Wells, M. Boers, N. Andersson, C. Hamel, A. C. Porter, P. Tugwell, D. Moher, and L. M. Bouter. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med. Res. Methodol.*, 7(1):10, 2007.

[42] B. J. Shea, B. C. Reeves, G. Wells, M. Thuku, C. Hamel, J. Moran, D. Moher, P. Tugwell, V. Welch, E. Kristjansson, and D. A. Henry. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, 358, 2017.

[43] J. A. C. Sterne, J. Savović, M. J. Page, R. G. Elbers, N. S. Blencowe, I. Boutron, C. J. Cates, H.-Y. Cheng, M. S. Corbett, S. M. Eldridge, J. R. Emberson, M. A. Hernán, S. Hopewell, A. Hróbjartsson, D. R. Junqueira, P. Jüni, J. J. Kirkham, T. Lasserson, T. Li, A. McAleenan, B. C. Reeves, S. Shepperd, I. Shrier, L. A. Stewart, K. Tilling, I. R. White, P. F. Whiting, and J. P. T. Higgins. Rob 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, 2019.

[44] D. F. Stroup, J. A. Berlin, S. C. Morton, I. Olkin, G. D. Williamson, D. Rennie, D. Moher, B. J. Becker, T. A. Sipe, S. B. Thacker, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*, 283(15):2008–2012, 2000.

[45] M. Turner, B. Kitchenham, P. Brereton, S. Charters, and D. Budgen. Does the technology acceptance model predict actual use? a systematic literature review. *Information and software technology*, 52(5):463–479, 2010.

[46] U. Wegewitz, B. Weikert, A. Fishta, A. Jacobs, and D. Pieper. Resuming the discussion of AMSTAR: What can (should) be made better? *BMC Medical Research Methodology*, 16(1):111, s12874–016–0183–6, Dec. 2016.

[47] R. J. Wieringa. Towards a unified checklist for empirical research in software engineering: first proposal. In *Proceedings of the 16th International Conference on Evaluation & Assessment in Software Engineering, EASE*, pages 161–165. IET, 2012.

[48] C. Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, pages 38:1–38:10. ACM, 2014.

[49] C. Wohlin. Writing for synthesis of evidence in empirical software engineering. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '14, pages 46:1–46:4, New York, NY, USA, 2014. ACM – IEEE, Association for Computing Machinery.

[50] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[51] L. Yang, H. Zhang, H. Shen, X. Huang, X. Zhou, G. Rong, and D. Shao. Quality assessment in systematic literature reviews: A software engineering perspective. *Inf. Softw. Technol.*, page 106397, 2020.

[52] C. Zapata. Integration of usability and agile methodologies: a systematic review. *Design, User Experience, and Usability: Design Discourse*, pages 368–378, 2015.

Appendix A: QAISER Instrument

Table 7: QAISER Instrument.

1. Did the authors of the review adequately justify the need for undertaking the review? [31, 37, 7]		
For Partial Yes:	For Yes:	
The authors of the review should have:	As for Partial Yes, plus the authors of the review should also have ALL of the following:	
☐ Identified existing related reviews on the topic, or explained that no related review exists	☐ Discussed related existing reviews on the topic, if any ☐ Established a scientific or practical need for their review [31]	☐ Yes ☐ Partial Yes ☐ No
Comments:		

2. Did the authors of the review establish a protocol prior to the conduct of the review? [42, 7, 31]		
For Partial Yes:	For Yes:	
The authors of the review confirm that a written protocol before the conduct of the review was established and is publicly available that provides details of the main elements of the systematic review process including the following:	As for Partial Yes, plus ALL of the following:	
☐ Appropriate review questions [42, 31, 37] ☐ Search process [42, 31] ☐ Study selection process [42, 31, 7] ☐ Data extraction process [31] ☐ Study quality assessment process (not relevant to most systematic mapping studies) [42, 31] ☐ An outline of the data synthesis plan [42, 31]	☐ Either authors of the review report that there no deviations from the protocol or any deviations are documented and justified [42, 7, 31]. ☐ The protocol should have been internally validated by piloting selection criteria, search strings, data extraction and synthesis processes [31].	☐ Yes ☐ Partial Yes ☐ No
Comments:		

3. Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?	
For Yes: the review should have ALL of the following:	
☐ Reported the inclusion and exclusion criteria [11, 31, 7] ☐ The criteria are aligned with the review questions ☐ Provided appropriate justifications for any restrictions used in the inclusion and exclusion criteria (e.g.,topic-related scoping restrictions, time-frame, language, study type, and peer reviewed works only) [31, 42]	☐ Yes ☐ No
Comments:	

4. Did the authors of the review use a comprehensive literature search strategy? [42, 31, 7, 11]
When database search is used as the main method

Table 7 continued from previous page

For Partial Yes:	For Yes:	
The review should have ALL of the following:	As for Partial Yes, plus the review should also have used:	
☐ An appropriate process for constructing the search strings including piloting [31, 7] ☐ The search date is appropriate, i.e., the review includes sufficiently new papers in relation to the paper submission date ☐ Search process validation based on using a known-set of papers [31, 7, 37] and providing an insightful discussion concerning why some papers are missing ☐ At least one relevant indexing database (e.g., Scopus) in combination with relevant publisher databases (e.g., IEEE and ACM) [31] ☐ Appropriately documented the search process (e.g., known-set, search strings, and search results) [31, 7, 4]	☐ At least one additional search method (e.g., snowballing, manual search, or use DBLP or Google Scholar of key researchers) [42, 31, 7]	☐ Yes ☐ Partial Yes ☐ No

When snowballing is used as the main method		
For Partial Yes:	For Yes:	
The review should have ALL of the following:	As for Partial Yes, plus the review should also have used at least ONE of the following:	
☐ Appropriately justified the use of snowballing as the main method [31] ☐ Selected an appropriate start/seed set. The selection process should be explained and justified [48] ☐ Search process validation based on using a known-set of papers [31, 7, 37] and providing an insightful discussion concerning why some papers are missing ☐ Performed an acceptable number of backward and forward snowballing iterations [48] ☐ Appropriately documented the search process (e.g., start/seed set, known-set, and search results) [31, 7, 4]	☐ At least one additional search method (e.g., manual search, or use DBLP or Google Scholar of key researchers) [48, 31] ☐ Snowballing iterations until no new papers were found [48]	☐ Yes ☐ Partial Yes ☐ No
Comments:		

5. Did the authors of the review use a reliable study selection process? [42, 31, 37]		
For Partial Yes:	For Yes:	
☐ At least two authors of the review selected a representative sample of eligible studies, achieved good agreement, and reported the agreement level, with the remainder selected by one review author [42, 31, 7].	☐ At least two authors of the review independently agreed on selection of eligible studies and reached consensus on which studies to include [42, 31, 37]. OR ☐ As with Partial Yes, with the additional requirement that at least two authors reviewed and achieved consensus about all excluded studies.	☐ Yes ☐ Partial Yes ☐ No
Comments:		

Table 7 continued from previous page

6. Did the authors of the review use a reliable data extraction process? [42, 31, 7, 37]

For Partial Yes:	For Yes:	
☐ At least two authors of the review extracted data from a sample of included studies, achieved good agreement, and reported the agreement level, with the remainder extracted by one review author [42, 31, 37].	☐ At least two authors of the review achieved consensus on which data to extract from the included studies [42, 31, 7, 37].	☐ Yes ☐ Partial Yes ☐ No

Comments:

7. Did the authors of the review discuss and justify the exclusion of the potentially relevant studies that were read in full text? [42, 31]

For Partial Yes:	For Yes:	
The review should have either ONE of the following:	As for Partial Yes, plus the review should also have the following:	
☐ Provided a list of all potentially relevant studies that were read in full text, but excluded from the review [42, 31]	☐ Justified the exclusion from the review of each potentially relevant study that was read in full text [42]	☐ Yes ☐ Partial Yes ☐ No

Comments:

8. Did the authors of the review provide sufficient primary studies' characteristics to interpret the results? [42, 31, 7, 11]

For Yes, the review should have described ALL of the following:

☐ Populations [42] ☐ Interventions, when applicable [42] ☐ Outcomes, when applicable [42] ☐ Study types [42] ☐ Study contexts [42]	☐ Yes ☐ No

Comments:

9. Did the authors of the review use an appropriate instrument for assessing the quality of primary studies that were included in the review? [42, 31]

For Yes, the review should have used appropriate instruments for different types of studies included in the review. An appropriate instrument would have questions related to ALL of the following [31]:

☐ *"The goals, research questions, hypotheses and outcome measures"* [31] ☐ *"The study design and the extent to which it is appropriate to the study type"* [31] ☐ *"Study data collection and analysis and the extent to which they are appropriate given the study design"* [31] ☐ *"Study findings, the strength of evidence supporting those findings, the extent to which the findings answer the research questions, and their value to practitioners and researchers"* [31]	☐ Yes ☐ No

Comments:

10. Did the authors of the review use a reliable quality assessment process? [31]

For Partial Yes:	For Yes:

Table 7 continued from previous page

☐ At least two authors of the review performed quality assessment of a representative sample of eligible studies and achieved good agreement, with the remainder performed by one review author [31].	☐ At least two authors of the review independently performed quality assessment of eligible studies and reached consensus about the quality levels or scores of the eligible studies[31].	☐ Yes ☐ Partial Yes ☐ No
Comments:		

11. Were the primary studies appropriately synthesized? [11, 31, 7, 8]	
For Yes, the review should have ALL of the following:	
☐ Selected an appropriate synthesis method given the review questions and extracted data [31, 7, 8] ☐ Applied the selected synthesis method appropriately ☐ Provided a clear trace linking the answers of review questions and interpretations to the data from the primary studies	☐ Yes ☐ No
Comments:	

12. Did the authors of the review investigate the impact of the quality of individual studies on the results of the review? [42, 31, 7]	
For Yes, either ONE of the following:	
☐ Included only high-quality studies [42, 31] OR ☐ The authors have discussed the impact of differences in the quality of individual studies on the results of the review [42, 31].	☐ Yes ☐ No
Comments:	

13. Did the authors of the review investigate the impact of primary studies' characteristics on the results of the review?	
For Yes, either ONE of the following:	
☐ There were no significant similarities or differences to warrant a separate analysis. OR ☐ The authors have discussed the impact of primary studies' characteristics on the results of the review [42, 31].	☐ Yes ☐ No
Comments:	

14. Did the authors of the review provide appropriate recommendations and conclusions from the review? [8]

For Partial Yes:	For Yes:	
The review should have the following:	As for Partial Yes, plus the recommendations and conclusions should also be:	
☐ Provided satisfactory recommendations and conclusions based on the review results	☐ Clearly traceable back to the review results ☐ Clearly targeting specific stakeholders ☐ Well aligned with the upfront motivation for undertaking the review, or are any deviations well explained ☐ Providing new valuable insights to the community	☐ Yes ☐ Partial Yes ☐ No

Table 7 continued from previous page

Comments:

15. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review? [42, 31]	
For Yes, either ONE of the following:	
☐ The authors reported no competing interests. OR ☐ The authors described their funding sources [42, 31] and how they managed any other potential conflicts of interest (e.g., how/who handled their own publications on the review topic while conducting the review) [42].	☐ Yes ☐ No

Comments:

Appendix B: QAISER Guidance Document

In this document, we provide further guidance to support a consistent interpretation of items in QAISER.

Item 1: Did the authors of the review adequately justify the need for undertaking the review?
A large number of SLRs are reported in software engineering every year. A review should be initiated on the basis of a practical or scientific need. The authors of the review should also extensively search for any existing reviews or mapping studies on the topic. The authors should only continue planning the review if there are no existing ones that are up to date on the specific area [31]. Mendes et al. [34] provide support to decide if a review should be updated.

To score 'Partial Yes', appraisers should check if the authors of the review have made a sufficiently extensive effort to identify related reviews on the topic. For example, a search using keywords for systematic secondary studies (like "systematic review", "systematic literature review", "systematic mapping study", or "systematic map") and topic specific keywords. For examples of such search please consult [1, 3].

To score 'Yes', in addition to the criterion under 'Partial Yes', appraisers should ensure that the authors of the review have established the need for undertaking the review. If there are existing reviews on the same topic, the authors need to establish the need by highlighting the gap in the existing reviews, and explaining how their review is going to fill the gap. In case there are no existing reviews on the topic, the authors explain why is it essential to aggregate the evidence on the topic.

The information about the need for review is typically described in the background or related work sections of the report.

Item 2: Did the authors of the review establish a protocol prior to the conduct of the review? To reduce the risk of bias, it is important that the authors of the review have developed and validated a written protocol before commencing the review.

To score 'Partial Yes', appraisers should first ensure that the authors of the review confirmed the establishment of a protocol before the conduct of the review. In addition, the protocol should be either accessible online and the review report describes where and how can it be accessed, or important parts of the protocol are included in the review report. Furthermore, the protocol should have documented appropriate review questions, the processes for search, study selection, data extraction, quality assessment and at least an outline for the data synthesis plan.

To rate 'Yes', appraisers should first check that all of the criteria under 'Partial Yes' have been met. In addition, the authors of the review should have clearly documented and justified any deviations from the protocol and discuss their impact on the study. Lastly, appraisers should also ensure that the protocol has been validated internally by pilot testing of different review processes (e.g., trial searches, selection, data extraction).

The above information about the protocol is typically described in the methodology section of the review report.

Item 3: Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions? A review should use documented selection criteria [7, 42, 31, 37, 11].

To score 'Yes', appraisers should ensure that the authors of the review have justified any restrictions, e.g., on research designs, the time frame of publication, and the type of publications imposed

in the selection process. Furthermore, the justification should also address the likely impact of the restrictions on the studied population and the generalization of the findings.

The selection criteria and the justifications for any restrictions are expected to be found in the methodology or limitations/threats to the validity section of the review report. Furthermore, some of the exclusion criteria may have been implemented in the search process.

Item 4: Did the authors of the review use a comprehensive literature search strategy? A comprehensive search strategy is important to maximize the coverage of the relevant literature. The authors of the review should provide a justification for using a particular search method (e.g., database or indexing service search or snowballing) as a primary method for searching the relevant literature.

To rate 'Partial Yes', appraisers should check the following in case of a database or indexing service search as the primary search method:

- The authors of the review have used an appropriate process for identifying the search terms, synonyms and constructing the search strings.
- The authors have included sufficiently new papers given the paper submission date. AMSTAR 2 recommends that the last search was conducted within the last 24 months. However, a different time may be more appropriate for the area of the review being assessed. The time may, for example, depend on the speed of the evolution of a specific area of research.
- The authors of the review have validated their search process by comparing their search results with a known-set of papers. The known-set of papers are the relevant papers that are already known to the authors of the review based on, for example, manual search or their knowledge of the review topic. The validation should, if papers are missing, include an insightful discussion concerning why some papers are missing.
- The authors of the review have used a combination of publisher databases and indexing services. IEEE and ACM are the most relevant publisher databases in software engineering, as they publish the most important and relevant conferences and journals in software engineering [31]. As a minimum, the authors should have used IEEE and ACM among the publisher databases and one indexing service (e.g. Scopus). Lastly, an important aspect of the application of the search strings to the selected databases is the timeliness of the search process. Appraisers should also ensure that the reported search is not outdated. While making this assessment of the timeliness of the search process, appraisers need to account for the time required to complete the remaining steps in the review process, writing the review report, and the peer-review process.
- The authors of the review have documented the search process. Appropriate documentation of the search process is important to ensure repeatability and transparency. The authors of the review should document: general and database specific search strings, total and database specific search results, search filters (e.g., years) used, date when the search strings were applied, known-set of papers used for validation, and validation measures (i.e. recall and precision) for details see [4].

To rate 'Yes', in addition to the above criteria, the authors of the review should have also used at least one additional search method (e.g., snowballing).

To rate 'Partial Yes', appraisers should check the following in case of a snowballing search as the primary search method:

- The authors of the review have used an appropriate process for identifying the seed set for starting the snowballing procedure. The way of identifying the seed set is well-documented and -motivated.
- The authors have included sufficiently new papers given the paper submission date. AMSTAR 2 recommends that the last search was conducted within the last 24 months. However, a different

time may be more appropriate for the area of the review being assessed. The time may, for example, depend on the speed of the evolution of a specific area of research.
- The authors of the review have validated their search process by comparing their search results with a known-set of papers. The known-set of papers are the relevant papers that are already known to the authors of the review based on, for example, manual search or their knowledge of the review topic. The validation is performed by computing recall and precision using the search results and the known set of relevant papers (for details, refer to [31]).
- The authors have iterated the snowballing procedure until no more papers are found.
- The authors of the review have documented the search process. Appropriate documentation of the search process is important to ensure repeatability and transparency. The authors of the review should document: identification of the seed set, the different iterations conducted, known-set of papers used for validation, and validation measures (i.e., recall and precision).

To rate 'Yes', in addition to the above criteria, the authors of the review should have also used at least one additional search method (e.g., manual search of key journals or conference proceedings, or use DBLP or Google Scholar of key researchers) or continuing snowballing iterations until no new papers are found).

Item 5. Did the authors of the review use a reliable study selection process? To reduce bias and the possibility of making mistakes, the crucial step of inclusion and exclusion of the papers should involve at least two reviewers [2, 42].

To rate 'Partial Yes', appraisers should check if at least two authors of the review selected a representative sample of eligible studies, achieved good agreement, and also reported the agreement level. While reporting the agreement level, the review report should also describe the sample size and the process adopted for achieving a good agreement level. Only after a Kappa score indicates that a good agreement has been achieved between at least two authors on a representative sample, a single author can proceed with the selection process for the remaining studies.

To rate 'Yes', appraisers should check that one of the two following processes are followed during study selection: 1) two authors of the review independently performed study selection on all eligible studies and reached consensus on which studies to include or exclude, 2) as with Partial Yes, i.e., two authors of the review selected a sample of eligible studies, achieved and reported good agreement level, with the remainder selected by one review author, but in that case all excluded studies must be reviewed by at least two authors of the review. A single reviewer should only proceed with the selection after a Kappa score indicating strong agreement between multiple authors of the review has been reached. However, even in this case, the excluded studies should be reviewed by at least one more author of the review - to ensure that decision to exclude any study is not made by a single author. The review would suffer from threats to validity if some studies were excluded by a single researcher. when it comes to including potentially irrelevant studies by a single author, the risk is mitigated by the fact that the irrelevance of such studies will become visible during the data extraction and synthesis processes. Therefore, it is important to ensure that all excluded studies are reviewed by at least two authors. Appraisers should also check that the rules for inclusion and exclusion, and how these rules were applied and how any differences between reviewers were resolved are described. Furthermore, the report should also report the number of papers remaining at each stage [8].

The information about the study selection process is expected to be described in the methodology and results sections of the review.

Item 6. Did the authors of the review use a reliable data extraction process? To ensure repeatability of the study and to avoid bias, it is important that the data extraction is not solely performed by a single researcher.

To rate 'Partial Yes', appraisers should check at least two authors extracted data from a sample of studies and have achieved a good agreement. The report should also include the agreement level level achieved, size of the sample, and the process used to achieve consensus on which data to extract.

To rate 'Yes', appraisers should ensure that data is extracted by at least two authors of the review. It is important to check that the review report provides a description of the process used to achieve consensus and shared understanding on which data to extract.

For SLRs that use qualitative synthesis methods, the data extraction process should be designed to minimize dependence on the viewpoint of a single member of the systematic review team and maximize the opportunities for team-based working and team decision making. Any tools used to assist data extraction and analysis should be identified and their contribution to data extraction process explained. The information of the data extraction process is generally described in the methodology section of the review report.

Item 7: Did the authors of the review discuss and justify the exclusion of the potentially relevant studies that were read in full text? This item refers to studies that were deemed relevant by authors of the review on a reading of the title and abstracts. However, after full-text reading, the authors concluded that the papers are not relevant to the current review. It is expected that the authors of the review should document such papers along with the reason for their exclusion. This will help increase confidence in the results, allow reflecting on the selection criteria used in the study, allow replications, and enable further research (for example, by leveraging on the filtered list of papers for a different analysis).

To rate 'Partial Yes', appraisers should see that the authors of the review have provided a list of such potentially relevant papers. Alternatively, the authors of the view should have reflected on the main reasons (e.g., papers that report no data for any of the review questions, papers excluded due to low quality) for excluding the papers that were read in full text.

In order to rate 'Yes', in addition to one of the alternatives for 'Partial Yes', justifications for excluding the potentially relevant papers should also be provided.

This documentation (i.e., a list of potentially relevant papers that were excluded after full-text reading and justifications for excluding them) can be made available in an appendix or as supplementary material for review online (along with other supporting material like the review's protocol).

Item 8: Did the authors of the review provide sufficient primary studies' characteristics to interpret the results?

The relevance and reliability of a systematic review depends, besides other factors, also on a number of factors related to the included studies such as its type (e.g., case study, survey, and experiment), context (real life or laboratory setting), participants (practitioners or students), and publication venue (e.g., a reputable conference/journal). The review report should describe adequate details about the characteristics of the included studies to inform the review readers about the kind of evidence that is used to draw conclusions. The concept of population in SE empirical studies is not limited to human subjects. In SE empirical studies, the focus may be on other items of interest - e.g., artifacts, issues or other events. Therefore, depending on the context, appraisers need to see what is a relevant population and whether or not authors of the review have included enough details about it in their review report.

To rate 'Yes', appraisers should ensure that the authors of the review have provided enough details about the population, interventions (when relevant), outcomes (when relevant), research designs and settings of the included studies.

These details may not be described at one place in a review report, and therefore could be challenging to find. Normally, part of this information is described in the start of the results section in a review report.

Item 9: Did the authors of the review use an appropriate instrument for assessing the quality of primary studies that were included in the review?

Due to several reasons, including the variety of research designs used in primary studies, reporting quality, use of inconsistent terminology, etc., quality assessment is a challenging task in software engineering systematic literature reviews [31, 28]. Several research-design specific checklists (e.g. for experimentation [50] and case study research [23]) and generic instruments (e.g. [47, 12]) have been proposed in literature. However, as concluded by Kitchenham [31], it is not feasible to use the same instrument to assess the quality of different types of studies.

To rate 'Yes', appraiser should ensure that the choice of the instruments used (whether an existing one or one formulated by the authors of the review) has been justified given the goals of the SLR and nature of included studies. Furthermore, the instrument used is expected to evaluate at least the research design, data collection, analysis reporting, and the strength of evidence given the stated goals of the primary study.

The information on the quality assessment of the primary studies is expected to be described in the methodology and results sections of the review report.

Item 10: Did the authors of the review use a reliable quality assessment process? Like Item 5 and 6, it is important that the quality assessment is not performed solely by a single author of the review.

To rate 'Partial Yes', appraisers should check if at least two authors have performed pilot quality assessment of a sample of the included studies to evaluate the objectivity of the quality assessment criteria and to develop a shared understanding of it.

To rate 'Yes', appraisers should see that at least two authors of the review independently performed the quality assessment of either all included studies or a sample of included studies (with the remaining performed by one review author) and achieved good consensus. The review report should also describe how differences were resolved in case of different quality scores.

The information about the quality assessment process is typically expected in the methodology section of the review report.

Item 11: Were the primary studies appropriately synthesized? Synthesis is one of the most important and also challenging parts of a systematic literature review. Without synthesis, the review would be of limited use.

In order to score 'Yes', appraisers need to see if the authors of the review have used and justified an appropriate method for synthesis. It may be the case that the authors of the review do not use the correct or appropriate name for the used synthesis method [14]. In that case, appraisers would have to carefully read the review report in order to make a decision on this item. The appraisers should further check if the selected synthesis method was appropriately applied and that there is a clear chain of evidence from the answers to the research questions to the data from the primary studies.

The information about the synthesis method and its output may be documented in a separate section. In some cases it may be described in the discussion section after the results section. It could

also be the case that the justification for selecting a specific synthesis method is described in the research methodology section, while the outputs of the synthesis step are described in a separate section.

Item 12: Did the authors of the review investigate the impact of the quality of individual studies on the results of the review? A review should take the quality of the individual studies into account when interpreting the results. This will increase the confidence in the findings and conclusions of the review.

To rate 'Yes', appraisers should see that either the review has excluded studies that do not meet the quality criteria defined in the study, or the authors have investigated and discussed the impact of differences in the quality of included studies on the results of the review. In the case of the first option - i.e., if the authors have only included high quality studies, the differences in the quality of study are not expected to be as high as requiring a further impact analysis. In the case of the second option - i.e., if the low quality studies are not excluded by the authors of the review, it is important that the authors investigate the impact of the quality of included studies on the results of the studies' synthesis by - for example - categorizing the results and analysis based on the quality of included studies. The information on using the quality of studies while interpreting the results is expected to be described in the discussion or analysis sections of the review report where results are further discussed/analyzed to draw conclusions.

Item 13: Did the authors of the review investigate the impact of primary studies' characteristics on the results of the review? There are many factors that can cause heterogeneity in the results of the included studies. It is important to analyze the causes of the heterogeneity in results, if any, while interpreting the results and drawing any conclusions. For example, it could be the variations in the contextual factors (e.g., student versus practitioners as subjects) that lead to differences in the results of different studies. Furthermore, quality scores or some specific quality criteria might also help in explaining the heterogeneity observed in the results [31]. This item is concerned with the use of study characteristics in Item 8.

In order to rate 'Yes', appraisers should see that the authors of the review have investigated the impact of study characteristics (including the number of studies) on the results of the review.

This discussion is likely to be found after the results section of the review report.

Item 14: Did the authors of the review provide appropriate recommendations and conclusions from the review? The usefulness of results of the review for the target stakeholders is critical to assess the relevance of the review. This item is a reflection on the aims as motivation for the review assessed in the first item of the instrument (i.e., item 1).

For 'Partial Yes' the review should have satisfactory recommendations and conclusions based on the review results.

For 'Yes', in addition providing satisfactory recommendations and conclusions, the recommendations and conclusions from the review shall also be traceable to the review results, clearly targeting specific stakeholders, well aligned with the motivation and provides new insights to the community.

Item 15: Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review? To ensure the reliability of a review, it is important that the authors of the review report their sources of funding and any other conflicts of interest. The disclosure of the sources of funding is quite obvious. However, identifying other types of conflicts of interest is not that straightforward.

For example, if the authors of the review have published on the topic of the review or have a vested interest in the outcome of the review, there is a potential for bias when selecting, analyzing and interpreting their own work and studies with competing alternatives.

It is encouraged that authors of the review should be experts in the topic area, so it is common that they have published extensively in the topic area. Thus, it is important that the authors of the review report their effort in identifying any conflicts of interest they have, which is relevant for the review. A mitigation strategy in this case is to establish a process and have it reviewed by independent researchers not participating in the literature review.

To rate 'Yes', the appraisers should ensure that the authors of the review have reported on the presence or absence of any conflicts of interest. In case there was some conflict of interest, the authors of the review should have described and justified the steps taken to mitigate the threat of bias in the results of the review.