When to update systematic literature reviews in software engineering

Emilia Mendes [a,*], Claes Wohlin [b], Katia Felizardo [c], Marcos Kalinowski [d]

[a] *Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden*
[b] *Department of Software Engineering, Blekinge Institute of Technology, Karlskrona, Sweden*
[c] *Department of Computing, Federal Technological University of Paraná, Cornélio Procópio, Brazil*
[d] *Department of Informatics, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil*

A R T I C L E   I N F O

A B S T R A C T

[*Context*] Systematic Literature Reviews (SLRs) have been adopted by the Software Engineering (SE) community for approximately 15 years to provide meaningful summaries of evidence on several topics. Many of these SLRs are now potentially outdated, and there are no systematic proposals on when to update SLRs in SE. [*Objective*] The goal of this paper is to provide recommendations on when to update SLRs in SE. [*Method*] We evaluated, using a three-step approach, a third-party decision framework (3PDF) employed in other fields, to decide whether SLRs need updating. First, we conducted a literature review of SLR updates in SE and contacted the authors to obtain their feedback relating to the usefulness of the 3PDF within the context of SLR updates in SE. Second, we used these authors' feedback to see whether the framework needed any adaptation; none was suggested. Third, we applied the 3PDF to the SLR updates identified in our literature review. [*Results*] The 3PDF showed that 14 of the 20 SLRs did not need updating. This supports the use of a decision support mechanism (such as the 3PDF) to help the SE community decide when to update SLRs. [*Conclusions*] We put forward that the 3PDF should be adopted by the SE community to keep relevant evidence up to date and to avoid wasting effort with unnecessary updates.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

In their call to arms paper, Kitchenham et al. (Kitchenham et al., 2004) argued for an Evidence-Based paradigm in Software Engineering (EBSE), to be particularly employed by *"researchers interested in empirical software engineering and practitioners faced with decisions about the adoption of new software engineering technologies."* Furthermore, they state that EBSE's goal is: *"to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision-making process regarding the development and maintenance of software"*, and also encourage the use of systematic literature reviews (SLRs) to obtain such current best evidence.

By mid-2008, 53 SLRs have been published in Software Engineering (SE) (20 found via a manual search of ten journals and four conferences (Kitchenham et al., 2009), and the remaining 33

via searches in six databases (Kitchenham et al., 2010)). Another 67 (Ferrari et al., 2018) SLRs were identified between January and December 2009; and finally an additional 316 SLRs between January 2010 and May 2016 (França et al., 2011), thus leading to a total of 436+ SLRs published in SE, within the period from January 2004 to May 2016. Furthermore, as detailed in Section 3, our database searches identified 20 updated SLRs. It provides evidence that many SLRs in SE are potentially outdated; thus affecting our current aggregated understanding of the state-of-the-art in those SLRs' research topics. Note that if we also consider potentially outdated primary studies, e.g. studies that refer to a particular technology that is not used any longer, the situation is even more challenging.

When new evidence is added as part of updating SLRs, different findings and conclusions from those reported initially may be identified. Therefore, updating SLRs may contribute to different purposes. For example, (a) providing a continuous update of the state-of-the-art on a research topic; and (b) identifying how that particular research topic is evolving (e.g. are the overall conclusions from previous SLRs stable throughout different SLR updates, or are they changing? If the latter, are these

---

* Corresponding author.
E-mail addresses: emilia.mendes@bth.se (E. Mendes), claes.wohlin@bth.se (C. Wohlin), katiascannavino@utfpr.edu.br (K. Felizardo), kalinowski@inf.puc-rio.br (M. Kalinowski).

changes significant enough to make previous SLRs' findings worth updating?).

How to keep SLRs current, more specifically concerning how and when to update SLRs, has also been a topic of discussion in other fields, such as medicine (e.g. (Garritty et al., 2010, Maplesden et al., 2015, Moher et al., 2008, Pattanittum et al., 2012)) and environmental sciences (e.g. (Zhang et al., 2008)). Although Cochrane has advocated in the past that SLRs in Medicine and other Healthcare specialities should be updated every two years (39), the guide, published in 2008 (Beecham et al., 2008), argued for a more detailed decision-making process, and provided a decision tool to guide such actions. In 2014, Cochrane organised a two-day workshop, attended by *"an international panel of authors, editors, clinicians, statisticians, information specialists, other methodologists, and guideline developers"*, aimed to identify *"a common approach to updating systematic reviews, drawing on existing strategies, research, and experience of people working in this area"* .[1] The results from this workshop were published in 2016 (Garner et al., 2016), where they also included a decision framework to guide the update (or not) of SLRs in a wide range of fields, not only Medicine.

Within the context of SE, SLRs also need updating; however, we have only found a short paper discussing this issue, and using the term SLR maintenance, rather than SLR update (Nepomuceno and Soares, 2018). They surveyed 28 authors of SLRs (not SLR updates) to identify the relevance to keep SLRs up-to-date and report that approximately 2/3 of those 28 respondents would like to keep their SLRs up-to-date. These results support the need for a decision framework or detailed guidelines to assess whether an SLR ought to be updated. Software Engineering can benefit from learning from other disciplines that are already doing this systematically. Therefore, in a similar way as Kitchenham (Kitchenham, 2004) and Kitchenham and Charters (Kitchenham and Charters, 2007), who used knowledge from more mature disciplines (in this regard, the knowledge relates to existing guidelines for conducting SLRs used by medical and social sciences researchers) to propose guidelines for conducting SLRs in SE, we have investigated existing decision frameworks and detailed guidelines in other more mature disciplines with respect to SLRs ((Garritty et al., 2010, Maplesden et al., 2015, Moher et al., 2008, Bayliss et al., 2016, Pattanittum et al., 2012, Hopewell et al., 2008)). We put forward that the decision framework proposed in (Garner et al., 2016) – 3PDF henceforth, can be used as a basis to guide the decision of when to update SLRs in SE. Section 2 introduces some existing decision frameworks in other disciplines, and details the 3PDF Framework, which is the one applied to assess SLR updates in SE (see Section 5). This is one of the main contributions of this paper.

Furthermore, we also argue that the definition of an SLR update, as suggested by Garner et al. (Garner et al., 2016), is broad enough to be adopted within the context of SE. Note that the definition resulted from discussions led by a panel of experts who carried out a variety of SLRs in a wide range of fields, not only restricted to Medicine and other Healthcare-related fields[1].

This panel's definition is as follows:

"an update of a systematic review [is defined] as a new edition of a published systematic review with changes that can include new data, new methods, or new analyses to the previous edition … An update asks a similar question with regard to the participants, intervention, comparisons, and outcomes (PICO) and has similar objectives; thus, it has similar inclusion criteria. These inclusion criteria can be modified in the light of developments within the topic area with new interventions, new standards, and new approaches. Updates will include a new search for potentially relevant studies and incorporate any eligible studies or data; and adjust the findings and conclusions as appropriate."

Within the context of this definition, new data represents new primary studies found via new searches – manual and/or in databases; new methods are, for example, to employ more detailed quality checklists to assess evidence, or to provide additional ways to aggregate evidence (e.g. thematic analysis), or to employ different ways to identify primary studies (e.g. database search vs snowballing[2]); finally, new analyses may include re-doing the data analysis based on, for example, more recent meta-analysis and research synthesis methods.

The specific goals and main contributions of this paper, which relate to when to update SLRs in SE, are fourfold: first, to identify existing policies, guidelines, recommendations relating to when to update SLRs to check whether they can be used within the context of SE; second, to carry out a literature review to identify SLR updates in SE; third, to carry out a survey aimed at the authors of these SE SLR updates; the survey's goal is to obtain these authors' feedback about using the policy/guideline/recommendation identified as part of our first research goal, as possible decision mechanism to support the decision about when/if to update an SLR in SE; and finally fourth, to apply the decision mechanism outcome of our third research goal to assess already published SE SLR updates (obtained as part of our second research goal). Based on our results, we provide recommendations on when to update SLRs in SE. These four goals are translated into four overall research questions (ORQ), addressed respectively in Sections 2 to 5:

ORQ1: What are (if applicable) the policies, guidelines, recommendations concerning when to update SLRs?
ORQ2: How have SLR updates been conducted in software engineering?
ORQ3: What is the perceived usefulness of the decision mechanism identified in ORQ1 to the authors of the SE SLR updates identified via ORQ2?
ORQ4: If using the decision mechanism (ORQ3), which SLR updates in SE (ORQ2) would be assessed as suitable for an update?

The remainder of this paper is organised as follows. In Section 2, we present some previous work in other fields relating to when to update SLRs and further describe the 3PDF framework. Section 3 reports a literature review of SLR updates in SE. In Section 4, we detail a survey employed to obtain feedback from the authors of the SLR updates presented in Section 3. Section 5 presents the use of the 3PDF to the SLR updates from Section 3. Section 6 contains a discussion on threats to validity, followed by an argumentation on research implications in Section 7. Finally, our conclusions and comments on future work are given in Section 8.

---

[1] Paul Garner, professor, Sally Hopewell, associate professor, Jackie Chandler, methods coordinator, Harriet MacLehose, senior editor, Elie A Akl, professor, Joseph Beyene, associate professor, Stephanie Chang, director, Rachel Churchill, professor, Karin Dearness, managing editor, Gordon Guyatt, professor, Carol Lefebvre, information consultant, Beth Liles, methodologist, Rachel Marshall, editor, Laura Martínez García, researcher, Chris Mavergames, head, Mona Nasser, clinical lecturer in evidence-based dentistry, Amir Qaseem, vice president and chair, Margaret Sampson, librarian, Karla Soares-Weiser, deputy editor in chief, Yemisi Takwoingi, senior research fellow in medical statistics, Lehana Thabane, director and professor, Marialena Trivella, statistician, Peter Tugwell, professor of medicine, epidemiology, and community medicine, Emma Welsh, managing editor, Ed C Wilson, senior research associate in health economics, Holger J Schünemann, professor.

---

[2] Note that within the context of this paper the term 'database search' represents a string-based automated search carried out on a database (e.g. SCOPUS) and/or Digital library (e.g. IEEEXplore, ACM). 'Backward snowballing' represents both a manual search over a paper's reference list and a string-based automated search on a database or digital library. Finally, 'forward snowballing' represents an automated search based on citation analysis.

Table 1
Terms derived from title, abstract and keywords found in studies from manual search.

Study	Conference	Main terms	Keywords
An extended systematic review of software process improvement in small and medium web companies	EASE 2011	Extended	Software Process Improvement
		Systematic Review	Systematic Review Small Web Companies Medium Web Companies
Motivation in software engineering: a systematic review update	EASE 2011	Update	Motivation
		Systematic Review Systematic Literature Review	Software Engineering Software Engineer Characteristics Personality Systematic Literature Review
Cross- vs. Within-Company Cost Estimation Studies Revisited: An Extended Systematic Review	EASE 2014	Systematic Review	Cost Estimation Models
		Extended Extending	Cross-company Data Within-company Data Estimation accuracy Systematic Review
Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study	ESEM 2015	Systematic Mapping	Replication
		Update	Empirical Software Engineering Systematic Mapping Study

2. Policies, guidelines or recommendations on when to update SLRs

2.1. Previous proposals

This sub-section introduces seven articles that were identified via searches carried out by the first author on Google Scholar and Scopus, using the following strings: "*When+and+how+to+update+systematic+reviews*", and "*update systematic reviews*". A total of 298 references were screened in Google Scholar and 35 in Scopus; seven and five references were selected, respectively. Note that after duplicates were removed, the final number of papers was seven. Also note that we found two different references to the same paper, which was published as an article (Moher et al., 2007) and also as a document in the Cochrane Database of Systematic Reviews 2008 (Moher et al., 2008). This search was not intended to be an SLR, and therefore we used only Google Scholar and Scopus. Except for one study, which was in the field of environmental management (Bayliss et al., 2016), all remaining studies were in the medical field.

Studies were arranged into three categories: i) systematic review about when and how to update SLRs; ii) techniques to identify new evidence related to previously published SLRs; and iii) decision mechanisms (factors or decision tree) to decide whether an SLR needs updating.

2.1.1. Systematic review about when and how to update SLRs

We found one article in this category, which had been published in two different places (Moher et al., 2007, Moher et al., 2008). Its authors looked for previous work (fully refereed papers and also gray literature) describing strategies and methods addressing when and/or how to update SLRs. They searched some databases - MEDLINE (from 1966 to 2005), PsycINFO, and also the Cochrane Methodology Register (Issue1, 2006), and manually searched the 2005 Cochrane Colloquium proceedings. They screened 2548 records and included 15 studies that characterized four strategies, one technique and two statistical methods describing when or how to update SLRs. We refer interested readers to Tables 1, 2, and 3 in (Moher et al., 2007). Their assessment shows several significant limitations throughout all proposals, and their overall conclusion is that "The identified strategies are not prag-

Table 2
Details on the search strategy.

Source	Studies Retrieved
Science Direct	722
Springer Link	241
IEEE Xplore	223
Scopus	640
ACM	199
Sub-Total	2025
Duplicates	356
TOTAL	1158

matic and have not been empirically tested. More concerted research efforts are needed to illuminate further knowledge gaps in the field of updating SRs and to ascertain the potential benefits of developing internationally harmonized, efficient, yet valid, ways of updating SRs.".

2.1.2. Techniques to identify new evidence related to previously published SLRs

Three studies assumed that the issue about when to update SLRs was solely driven by the existence of more recent evidence relating to an SLR (Sampson et al., 2008, Aug, Chung et al., 2012 Jun, Pattanittum et al., 2012). Two studies were only based on the results from manual database searches (Sampson et al., 2008, Aug, Chung et al., 2012 Jun); the third study used a combination of manual searches and statistical techniques that assessed whether new evidence would change the conclusions from meta-analyses based upon evidence included in the perceived outdated SLRs (Pattanittum et al., 2012). In this latter case, any SLRs candidate for update should also include meta-analyses.

2.1.3. Decision mechanisms (factors or decision tree) to decide whether an slr needs updating

The first study to present a decision mechanism to decide whether an SLR was eligible for updating was carried out by Takwoingi et al. (Takwoingi et al., 2013 Dec 13). They proposed a decision tool, which in fact can be seen as a precursor to Garner et al.'s decision tool (Garner et al., 2016) (presented in detail in sub-Section 2.2). This decision tool supports the judgement as

Table 3
Studies analysed during Phase 2.

ID	Study Title	Study Selection	
		Initial	Final
S1	A Survey on Software Release Planning Models	I	I
S2	A Systematic Literature Review of Techniques and Metrics to Reduce the Cost of Mutation Testing	I	I
S3	A systematic literature review on cloud computing adoption and migration	I	I
S4	An extended systematic literature review on provision of evidence for safety certification	I	I
S5	An extended systematic review of software process improvement in small and medium web companies	I	I
S6	An update to the SLR of empirical evidence of the impacts and outcomes of computer games and serious games	I	I
S7	Cloud service evaluation method-based Multi-Criteria Decision-Making: A systematic literature review	I	I
S8	Critical success factors for ERP system implementation projects: An update of literature reviews	I	I
S9	Cross vs. within-company cost estimation studies revisited: An extended systematic review	I	I
S10	Effort Estimation in Agile Software Development: An Updated Review	I	I
S11	Knowledge management practices in GSD – a systematic literature review update	I	I
S12	Maintainability prediction of relational database-driven applications: A systematic review	I	I
S13	Motivation in software engineering: a systematic review update	I	I
S14	Open-Source Development Tools for Domain-Specific Modeling: Results from a SLR	I	I
S15	Process simulation for software engineering education	I	I
S16	Revisiting software ecosystems Research: A longitudinal literature study	I	I
S17	State-of-the-Art: A Systematic Literature Review on Agile Information Systems Development	I	I
S18	Systematic literature review on agile practices in global software development	I	I
S19	Updating a systematic review about selection of software requirements elicitation techniques	I	I
S20	What is the Further Evidence about UML? - A Systematic Literature Review	I	I
S21	How personality traits are interrelated with team climate and team performance in SE? A preliminary study	I	E
S22	Is There a Place for Qualitative Studies when Identifying Effort Predictors? A Case in Web Effort Estimation	I	E
S23	Obtaining well-founded practices about elicitation techniques by means of an update of a previous SR	I	E
S24	SOA Reuse: Systematic Literature Review Updating and Research Directions	I	E
S25	Software process simulation modeling: An extended systematic review	I	E
S26	Software process simulation modeling: preliminary results from an updated systematic review	I	E
S27	Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study	I	E

to whether an SLR is qualified for updating via three consecutive steps (detailed in Fig. 2 in their paper):

"Step 1: Is the clinical question already answered by the available evidence or is the clinical question deemed no longer relevant?

Step 2: Are there any new factors relevant to the existing review? (e.g. additional information from existing included studies, previously missing data)

Step 3: Are there new studies?"

The tool assumes that all SLRs should both relate to clinical trials and include a meta-analysis. Furthermore, SLRs are judged as not in need for updating when there are new factors to consider but when applying the new factor(s) there is no new evidence and/or new conclusions. Overall, this decision tool is much more limiting when compared to Garner et al.'s decision tool, which, amongst other things, was proposed for use within and outside the medical field. Note that the paper's two first authors were also part of the team that participated in Garner et al.'s study and workshop, which is also detailed in Sub-Section 2.2.

The second study, by Bayliss et al. (Bayliss et al., 2016), proposes factors that are believed to affect the decision as to whether an SLR update is needed. The set of factors (listed below) is not arranged as part of a decision tool but simply listed. Furthermore, the proposal is much simpler than the ones by Takwoingi et al. (Takwoingi et al., 2013 Dec 13) and Garner et al. (Garner et al., 2016) and focuses on whether new evidence on the SLR's topic has been published. The factors are as follows:

a The SLR's topic is still relevant to policy and practice, and some preliminary searches suggest that there are new studies suitable for inclusion.

b The SLR's topic is relatively new, and the original SLR included limited data. An update should be done if it is acknowledged that such an update would provide valuable additional information.

c "Where large volumes of information have been published over a short timescale, for example for topical issues (e.g. pollinators) or broad areas (e.g. climate change), especially where evidence is contentious or where trends in research methods, study subjects, research groups or dogma have demonstrated a shift over time—new evidence may be more likely to contradict old."

d If large influential studies were published and may affect the original SLR's conclusions.

The third study, by Garner et al. (Garner et al., 2016), details a decision tool, arranged as a decision tree, which considers different aspects relating to the suitability for an SLR to be updated (not only the existence of new evidence). The proposal, to be detailed in the next sub-section, is more comprehensive than all the other suggestions abovementioned, and also aims to support decisions in a wider range of fields outside the medical field.

2.2. Decision framework proposed by garner et al.

Garner et al.'s (Garner et al., 2016) decision framework was selected within the context of this research for several reasons. One of these reasons has already been mentioned in Sub-Section 2.1 and relates to including decision points that go beyond the existence of newly published evidence on the SLR's topic. The framework, which is presented in Fig. 1, also supports the updating of SLRs based upon other considerations such as the use of new relevant methods. Another reason for choosing this framework is that it is the only study amongst those introduced in Sub-Section 2.1 that targets an audience both within and outside the medical field. Therefore, we believe that it is generic enough so to also be applicable within the SE context.

Furthermore, Garner et al.'s decision framework resulted from a collaboration between different stakeholders aiming to put forward as decision support to update SLRs in different areas, by drawing on existing *"strategies, research, and experience of people working in*
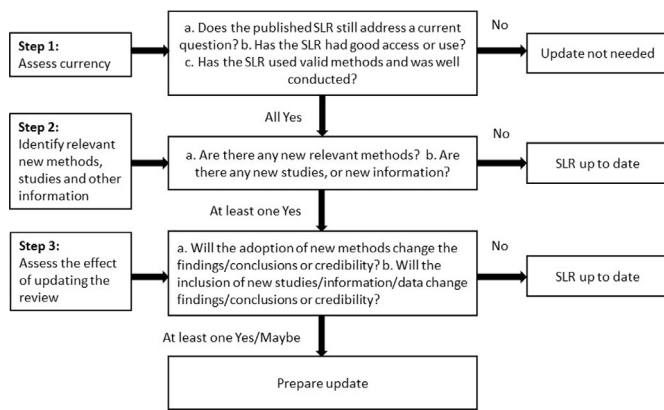
Fig. 1. Decision framework to assess SLRs for updating (Garner et al., 2016).

this area" (Garner et al., 2016). Therefore, it is an improvement over the tools that had been previously suggested by Cochrane and others. Note that Section 4 describes a survey in which feedback about the use of this framework within SE is discussed. Moreover, we also evaluate this framework by applying it to twenty SLR updates in SE.

The decision framework includes three steps (See Fig. 1), to be applied sequentially, as follows.

Step 1 focuses on assessing how current/actual the SLR is by looking at its topic's relevance for research and practice. This includes whether the published SLR has had an impact on research and/or practice (using metrics such as citations via sites such as Google Scholar), and finally whether the SLR was carried out properly and using a sound methodology (e.g. clear inclusion/exclusion criteria and quality assessment). As shown in Fig. 1, if the answer to at least one of Step 1's questions is 'NO', there is no arguable need for updating this SLR. Note that in the particular case where questions a. and b. are 'YES' and question c. 'NO', an update of the original SLR is not warranted; however the original SLR should in fact be redone given that the method originally used is not viewed as valid. While being redone, it can also widen its search so to include more up to date primary studies. The above mentioned framework does not target this situation, given that it is not an update, but a new SLR. Possible answers are 'YES' and 'NO'.

Step 2 focuses on identifying if there are any new methods proposed and/or new studies published after the SLR's publication. Examples of new methods can be new ways to search for evidence and to do the synthesis of evidence, the inclusion of other types of study designs (e.g. qualitative research), new statistical techniques or improvements to existing techniques used for meta-analysis. Step 2.b relates specifically to whether new additional studies were found. In this step, as long as at least one question is affirmative, one can proceed to the next step – Step 3. Possible answers are 'YES' and 'NO'.

Finally, **Step 3** targets at assessing whether the adoption of new methods and/or new studies may affect the conclusions when compared to the conclusions from the original SLR, and/or the original SLR's credibility. Note that concerning 3.b, there is a range of different aspects to consider. For example, conclusions may change due to adding new primary studies to the original SLR. These would be identified by looking at the results from database searches/snowballing, based on reading titles and abstracts. Note that such a task will most likely be positively impacted whenever prospective studies use structured abstracts. Other examples would be changes made to the inclusion criteria or quality checklist which, when reapplied to the studies in the original SLR,

lead to changes in the conclusions. The same applies here as in Step 2, i.e., as long as at least one question is affirmative, it is recommended to proceed to the next step. Possible answers are 'YES/MAYBE' and 'NO'.

At the end of Step 3, it is clear that an SLR needs to be updated. Suggestions on how to update such SLRs are provided in Section 5, and details on how to carry out searches are given in Mendes et al. (Mendes et al., 2019).

3. Searching for existing slr updates in se

This section details a Systematic Literature Review (SLR) carried out to identify SLR updates in Software Engineering (SE). The reasons for carrying out this SLR were three-fold: first, to identify a sample of SLRs updates in SE. We use the word sample because no SLR that has as one of its inclusion criterion "the study is written in English" can claim full coverage of studies. Second, to identify SE SLR update authors. These authors are invited later to participate in a survey aimed at obtaining their feedback in relation to: i) the adoption of the 3PDF framework within the context of SE SLR updates; and ii) the framework itself. This survey is detailed in Section 4; and third, to apply the 3PDF framework (and any recommended changes resulting from the survey investigation) to each of the SE SLR updates presented in this section (this is detailed in Section 5).

Research Questions

There are seven Research Questions (RQs), in addition to the four overall research questions (ORQ), we are interested in, which are directly related to the three Steps used in the 3PDF framework when deciding whether an SLR needs to be updated. We have matched the RQs with the framework steps because these drove both the data extraction phase and the results presented in this section.

The RQs are as follows:

- **RQ1** (relates to Step 1.a. in the 3PDF): Does the published SLR still address a current question?
- **RQ2** (relates to Step 1.b. in the 3PDF): Has the SLR had good access or use?
- **RQ3** (relates to Step 1.c. in the 3PDF): Has the SLR used valid methods and was well-conducted?
- **RQ4** (relates to Step 2.a. in the 3PDF): Are there any new relevant methods?
- **RQ5** (relates to Step 2.b. in the 3PDF): Are there any new studies, or new information?
- **RQ6** (relates to Step 3.a. in the 3PDF): Will the adoption of new methods change the findings, conclusions or credibility?
- **RQ7** (relates to Step 3.b. in the 3PDF): Will the inclusion of new studies/information/data change findings, conclusions or credibility?

Inclusion and Exclusion Criteria

We used as inclusion criteria (IC):

- **IC1**: The study is an updated SLR; and
- **IC2**: Both the original secondary study and its update are available, and provide the data that is needed for comparison, e.g., the search string used in both versions or appropriate descriptions of the searches.
- **IC3**: The study is a fully refereed paper at either a conference or journal, and.
- **IC4**: The study is written in English.

Our exclusion criteria (EC) are listed next, and items that are self-explanatory are not detailed further:

- **EC1**: The study is duplicated (e.g., returned by different databases),

- **EC2**: The original study is not an SLR. Here both the original and the update must be SLRs, so to provide a basis for comparison (e.g. search string used),
- **EC3**: The updated study is a systematic mapping study or a survey. The 3PDF is a decision tool to be applied to SLR updates only, so within the context of this research we maintained the same focus,
- **EC4**: The updated study is a previous version of a more recent update. Here we applied a similar approach as to the one used in SLRs with regard to primary studies: whenever there are different versions of the same primary study, the one to use is the most complete description. Within our context, the latest update would be assumed to be the most complete,
- **EC5**: The study is not a fully refereed paper at either a conference or journal. Here we are neither interested in gray literature nor very brief descriptions such as those in short papers, and.
- **EC6**: The study is written in a language other than English.

Identifying Search Terms to Use

The first step, prior to carrying out the database searches on several different databases, was to identify the terms to use to build our search string. To do so, the third author carried out a manual search in the proceedings of two conferences – ESEM (International Symposium on Empirical Software Engineering and Measurement) and EASE (International Conference on Evaluation and Assessment in Software Engineering), for the period 2006 to 2018, based upon titles, keywords and abstracts. These conferences were selected for having a clear focus on Evidence-based Software Engineering, and for including secondary studies as some of the topics of interest. We limited the manual search to these two conferences since the objective was to identify the terms for the database search, and not to find studies to include. A total of 794 studies were identified (EASE: 233 and ESEM: 561), and four SE updates were found (three from EASE and one from ESEM proceedings, respectively). The selection was discussed with the first author, and a total agreement was reached. Note that these four studies were also retrieved via our database searches (see Table 3 - S5, S9, S13, and S27), the purpose of our manual search was solely to identify the terms for the database search, and not to find studies to include. All the included studies come from the database searches. During the manual search, terms and synonyms from titles, abstracts and keywords were derived to use later in the database search. The main terms extracted from the manually included studies are shown in Table 1.

The terms in Table 1 led to the following search string:

---

*("systematic review" **OR** "systematic literature review" **OR** "systematic mapping")*
***AND** (update **OR** extended **OR** extending)*

We used the Boolean **OR** to incorporate synonyms or alternative words or abbreviations and finally, use the Boolean **AND** to link the major terms. The final string was:

---

*("systematic literature review" **OR** "systematic review" **OR** "systematic study" **OR** "systematic analysis" **OR** "mapping study" **OR** "systematic mapping" **OR** "mapping review") **AND** (update **OR** updating **OR** updated **OR** extended **OR** extending **OR** extend **OR** extension)*

Note that we used the word mapping to also cater for any possible paper(s) that were genuine SLRs but perhaps labelled themselves as mapping studies.

Databases used for searches

The databases selected for searching were: ACM Digital Library; IEEE Xplore; ScienceDirect; Scopus and Springer Link. These databases were selected for being databases recommended by previous SLRs as the adequate online databases to use (e.g. Maplesden et al. (Maplesden et al., 2015), and Zhang et al. (Zhang et al., 2011)).

		Evaluator B			
		0	**1**		
Evaluator A	**0**	0	3	3	12%
	1	1	23	24	92%
		0	33	26	
		4%	100%		

Fig. 2. Confusion matrix for Phase 2.

Note that there are other databases (e.g. EI compendex, Wiley Interscience (Wiley Online), Inspec and Kluwer) that were not selected due to the high degree of overlap between their results and those for the selected databases (Maplesden et al., 2015).

All 1158 potentially relevant primary studies were stored using the JabRef software, an open-source bibliography reference manager. We used the export feature available in many electronic databases to automatically export the details of all potential primary studies (i.e., title, author(s), abstract, keywords, year of publication and the name of the data source) to JabRef. Information from databases that did not support exporting to JabRef was manually recorded by the third author. The third author was also the one dealing with duplicates. Finally, the selection of studies was carried out using two phases, as follows.

- **Phase 1:** the first and third authors independently applied the inclusion & exclusion criteria to screen 1065 and 1158 studies, respectively. The strategy used by both was different. The first author clustered, with the JabRef tool support, studies using searches based on terms, e.g., studies that contained the *"systematic review"* term in its title. Subsequently, in this subset of studies, she identified studies that contained, e.g., terms such as *"update"* or *"extended"*. In each subset of studies, the selection criteria were applied by reading the titles and abstracts. The third author sequentially applied the inclusion/exclusion criteria reading the titles/abstracts of all studies organized in the JabRef tool. The individual results were compared and resolved by consensus; a subset of 27 studies was chosen for Phase 2 (detailed next).
- **Phase 2:** the full text of the 27 primary studies included in the preliminary selection (see Table 3) was evaluated independently by the first and third authors to decide whether the study should be included or excluded. The few disagreements regarding the eligibility of studies were also resolved by consensus. The confusion matrix corresponding to the selection is given in Fig. 2. It clearly shows that there was a very high overlap between the two authors as to which papers were judged as relevant (23 studies). This meant that Cohen's kappa turned out to be negative (-0.059), suggesting that the two authors' agreement strength is worse than what would be expected by chance alone. A possible explanation for the results could be the detailed level of screening applied in Phase 1, leaving for Phase 2 just a small number of false positives. A total of 20 studies were selected, and seven excluded. The reasons for exclusions are shown in Table 4, and these excluded studies' references are listed in Appendix A.

A summary of the entire process, from selecting search terms to identifying the SE SLR updates, is shown in Fig. 3.

Note that in three cases (S9, S15 and S18 in Table 5), the original SLR (OS) had already been revisited at least twice. Under such circumstances, we chose to compare the latest SLR update with its previous most recent version (e.g. S15 is compared to S15_OSc). The only exception is S9, which is compared to S9_OSa (instead of S9_OSb). This was done because, although S9_OSa was not the most recent version before S9, it was the version that presented results for the same subset of research questions discussed in S9.

Table 4
Studies excluded during Phase 2 and reasons for the removal.

ID/ Ref	Study's Title	Reason for exclusion
S21/ (Soomro et al., 2015)	How personality traits are interrelated with team climate and team performance in SE? A preliminary study	EC3: It is not an SLR updated. It is a data survey
S22/ (Matos et al., 2014)	Is There a Place for Qualitative Studies when Identifying Effort Predictors? A Case in Web Effort Estimation	EC2: The original study is not an SLR. It is a traditional review
S23/ (O. Dieste et al., 2008)	Obtaining well-founded practices about elicitation techniques by means of an update of a previous SR	EC4: It is the oldest version of study S19 (O. Dieste et al., 2008)
S24/ (Oliveira et al., 2018)	SOA Reuse: Systematic Literature Review Updating and Research Directions	EC2: The original study is not an SLR. It is an MS
S25/ (Zhang et al., 2010)	Software process simulation modeling: An extended systematic review	EC4: It is one of the original studies of S15 (Jiang et al., 2015)
S26/ (Gao et al., 2014)	Software process simulation modeling: preliminary results from an updated systematic review	EC4: It is one of the original studies of S15 (Jiang et al., 2015)
S27/ (Bezerra et al., 2015)	Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study	EC2: The original study is not an SLR. It is an MS
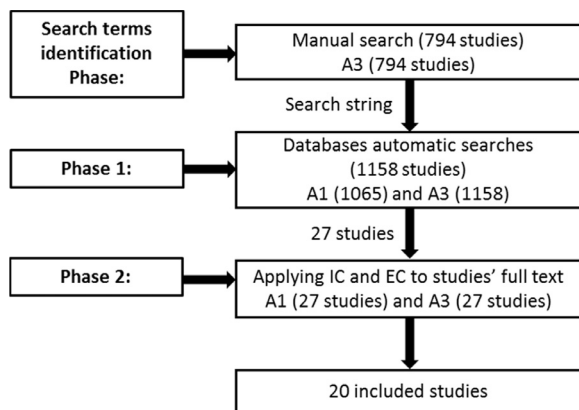
Fig. 3. Summary of the process used to identify search terms and SLR updates.

In fact, S9_OSa and S9_OSb complement each other and represent the same SLR.

Answering the Research Questions

RQ1 - Does the published SLR still address a current question?

Table 5 provides some details about the SE SLR updates that were included herein. The oldest SE SLR update dates back to 2008 (S19), followed by two published in 2011 (S5 and S13). Perhaps one could argue that topics published many years ago would not be meaningful to SE anymore. However, the chosen topics – requirements elicitation techniques, motivation in SE, and software process improvement, are still relevant nowadays and are explicit topics of some SE conferences and journals. Furthermore, all the other SLR updates were published quite recently. Thus, our joint expert judgement is that all the 20 SE SLR updates presented in this section addressed current important SE questions.

RQ2 - Has the SLR had good access or use?

To answer this RQ, we used as cut-off point the same average citation value that has been documented by Garousi and Fernandes (Garousi and Fernandes, 2016). They gathered data from 71,668 SE papers in the pool indexed by Scopus, to identify the top-100 most cited SE papers. They documented that, using the 71,668 papers as basis, the average citation value was 6.82 per paper. Given the large sample size, we believe that it would be reasonable to adopt within the context of our work the same average citation value as a cut-off. Note that all the updates were considered (e.g. S15 had three updates; S9 and S18 had two each). SLR updates with a yearly average equal or above 6.82 were judged as having had good access or use. The average yearly citation is shown in bold in

Table 5. Note that some citations in Table 5 are crossed as those SLRs, despite being shown, are not considered within the context of the work presented herein. Only the most recent previous SLRs are considered. Therefore, we have eleven previous SLRs that were judged as presenting good access or use (S1, S6, S9, S10, S12, S13, S16, S17, S18, S19, and S20).

With regard to using the average citation as a surrogate measure for relevance, one of the reasons for doing so was to comply with the same suggested measures proposed in the 3PDF, where citation count and/or website downloads (e.g. as in ResearchGate) were recommended as surrogates for relevance. Another reason was that citation count is a measure that has been used in quite a number of bibliometrics studies in SE, where citations are employed as a measure of research quality and relevance. One example of such use is provided in the same paper abovementioned, by Garousi and Fernandes (Garousi and Fernandes, 2016), where they employ citation count as a surrogate for relevance when identifying the top 100 most highly cited SE papers. They also provide a detailed list with numerous bibliometrics studies in SE that have assumed citation count as a surrogate for relevance. Finally, there are also numerous bibliometric studies in other disciplines that also use citation count as surrogate for relevance (Bornmann, 2014).

RQ3 - Has the SLR used valid methods and was it well-conducted?

To answer this RQ, the full text of the SLR updates was used to understand how they had been conducted, and to populate Tables 6 to 8.

Table 6 details the SLR method adopted by the 20 SLR updates. The updated versions' protocols were strongly based on the previous version, and we targeted the search strategy, data extraction, and the data synthesis employed to identify whether there were changes between SLR versions. These categories are also the ones suggested in Garner et al. (Garner et al., 2016).

As can be seen in Table 6, fifteen SLRs (S2, S3, S5, S6, S8, S9, S10, S11, S12, S13, S14, S15, S16, S18 and S19) answered the same RQs of the previous reviews; four (S4, S7, S17 and S20) added a new question or sub-question, and one (S1) discarded some RQs used in the previous version. Furthermore, fifteen SLRs (S2, S3, S5, S8, S9, S10, S11, S12, S13, S14, S16, S17, S18, S19 and S20) reused the same inclusion/exclusion criteria from the original studies; three (S4, S6 and S15) included an additional exclusion criterion based on publication date to exclude the studies included in the previous version; and two studies (S1 and S7) included additional inclusion/exclusion criteria.

Concerning the search strategy, it can be seen that twelve SLRs (S2, S3, S5, S7, S8, S11, S12, S13, S14, S16, S18 and S20) used the

Table 5

Details on types of search and where searches were done.

Updated SLR					Original/Previous SLR			
ID	Research Topic	Year	Search span	Timespan bt. SLRs' publication dates	Original SLR (OS)	Search span	Year	Cited by Google Scholar/avg. Citations per year
S1 (Ameller et al., 2016)	Software Release Planning	2017	2009 – 2015	7 years	S1_OS (Svahnberg et al., 2010)	open – 2008	2010	158/17.6
S2 (Pizzoleto et al., 2019)	Mutation Testing	2019	2018 – 2019	1 year	S2_OS (Ferrari et al., 2018)	1988 – 2018	2018	4/4
S3 (de Paula and de Carneiro, 2016)	Cloud Computing	2016	2015 – 2016	0 year	S3_OS (de Paula and Carneiro, 2016)	2005 – 2015	2016	1/0.33
S4 (Nair et al., 2014)	Safety Certification	2014	open – 2012	1 year	S4_OS (Nair et al., 2013)	open – 2012	2013	25/4.17
S5 (Sulayman and Mendes, 2011)	Software Process Improvement (SPI)	2011	2008 – 2010	2 years	S5_OS (Zhang et al., 2008)	open – 2007	2009	51/5.1
S6 (Boyle et al., 2016)	Computer Games and Serious Games	2016	2009 – 2014	4 years	S6_OS (Connolly et al., 2012)	2004 – 2009	2012	878/125.43
S7 (Alabool et al., 2018)	Cloud Computing	2018	2006 – 2016	5 years	S7_OS (Alabool and Mahmood, 2013)	2007 – 2012	2013	10/1.7
S8 (Leyh and Sander, 2015)	Enterprise Resource Planning system	2015	2010 – 2013	3 years	S8_OS (Leyh and C., 2012)**	open – 2010	2012	21/5.25
S9 (Mendes et al., 2014)	Cost Estimation	2014	2006 – 2013	7 years	S9_OSa (Kitchenham, 2004)	1990 – 2006	2006	~~111/8.54~~
					S9_OSb (Kitchenham et al., 2007)	1990 – 2006	2007	338/28.17
S10 (Dantas et al., 2018)	Effort Estimation	2018	2014 – 2017	4 years	S10_OS (Usman et al., 2014)	2001 – 2013	2014	113/22.6
S11 (Beninca et al., 2015)	Knowledge Management (GSD*)	2015	2012 – 2014	3 years	S11_OS (Arshad et al., 2012)	2008 – 2012	2012	00/0
S12 (Riaz, 2012)	Maintainability Prediction	2012	2008 – 2010	3 years	S12_OS (Riaz et al., 2009)	1985 – 2008	2009	214/21.4
S13 (França et al., 2011)	Motivation in SE	2011	2006 – 2010	3 years	S13_OS (Beecham et al., 2008)	1980 – 2006	2008	458/41.64
S14 (Hoisl and Sobernig, 2016)	Open-Source Development	2016	2005 – 2012	0	S14_OS (Sobernig et al., 2016)	2005 – 2012	2016	13/4.33
S15 (Jiang et al., 2015)	Process Simulation	2015	open – 2013	1 year	S15_OSa (Zhang et al., 2008)	1998 – 2007	2008	~~94/8.55~~
					S15_OSb (Zhang et al., 2011)	1998 – 2007	2010	~~53/5.89~~
					S15_OSc (Gao et al., 2014)	2008 – 2012	2014	04/0.8
S16 (Manikas, 2016)	Software Ecosystem	2016	2007 – 2014	3 years	S16_OS (Manikas and Hansen, 2002)	2007 – 2012	2013	116/19.33
S17 (Hummel, 2014)	Agile Software Development	2014	2010 – 2013	2 years	S17_OS (Dingsøyr et al., 2012)	2001 – 2010	2012	797/113.86
S18 (Vallona et al., 2018)	Agile practices	2018	2010 – 2016	8 years	S18_OSa (Jalali and Wohlin, 2012)	1999–2009	2011	~~133/16.63~~
					S18_OSb (Jalali et al., 2010)	1999–2009	2010	156/17.33
S19 (Sulayman and Mendes, 2011)	Requirements Elicitation Techniques	2008	No new search	3 years	S19_OS (Dieste and Juristo, 2010)	open – 2005	2011	141/17.63
S20 (Guo et al., 2017)	UML	2017	2009 – 2016	6 years	S21_OS (Budgen et al., 2011)	open – 2008	2011	119/14.88

* GSD: Global Software Development.

** Both reviews (S8 and S8_OS) were published in a single paper (Leyh and Sander, 2015).

Table 6
SLR method adopted during the updates.

Research Question (RQ)	
Answered the same RQ(s) of OS	S2, S3, S5, S6, S8, S9, S10, S11, S12, S13, S14, S15, S16, S18, S19
Added a new RQ or sub-question	S4, S7, S17, S20
Discarded some RQ(s) of OS	S1
I/E Criteria	
Used the same I/E criteria of OS	S2, S3, S5, S8, S9, S10, S11, S12, S13, S14, S16, S17, S18, S19, S20
Included new I/E criterion based on publication date	S4, S6, S15
Included new I/E criteria	S1, S7
Search strategy	
Used the same search strategy of OS	S2, S3, S5, S7, S8, S11, S12, S13, S14, S16, S18, S20
Used snowballing only in OS	S19
Used snowballing only in the updated study	S1, S10
Used a subset of database(s) used in OS	S6
Used additional database(s) to those used in OS	S4, S9
Changed the search string used in OS	S15
Changed the search strategy used in OS	S17
Quality assessment (QA)	
Used the same quality criteria of OS	S3, S4, S5, S6, S9, S10, S11, S12, S13, S14, S15, S16
Defined additional criteria for quality assessment	S18
Did not apply QA to the updated version	S1, S20
Did not apply QA to both versions	S2, S7, S8, S17, S19
Data extraction	
Used the same data extraction form of OS	S2, S3, S4, S5, S6, S8, S9, S10, S11, S12, S13, S14, S15, S16, S18, S19, S20
Used partially the data extraction form of OS	S1, S7, S17
Data synthesis	
Integrated data in OS' results	S2, S3, S5, S11
Compare results with those from OS	S1, S4, S6, S7, S8, S9, S10, S12, S13, S15, S16, S17, S18, S19, S20
Included different results from OS	S14
Team	
The same team of OS	S3, S5, S8
Involved at least one member of OS' team	S2, S4, S6, S7, S9, S12, S14, S15, S16, S19
Created a new team	S1, S10, S11, S13, S17, S18, S20

Table 7
Details of the search strategy adopted during the updates.

ID	Search strategy		
	Type of Search	Search string	Databases (dbs)
S1	SN* Back/Forward	DS was not used	DS was not used
S2	DS + SN* Back	The same search string as in the OS	The same dbs as in the OS
S3	DS + MS	The same search string as in the OS	The same dbs as in the OS
S4	DS + MS	The same search string as in the OS	New dbs were used *
S5	DS	Based on keywords of the OS*	The same dbs as in the OS
S6	DS	The same search string as in the OS	A subset of those dbs used in the OS
S7	DS	Based on keywords of the OS	The same dbs as in the OS
S8	DS	The same search string as in the OS	The same dbs as in the OS
S9	DS + SN* Back/Forward	The same search string as in the OS	New dbs were used *
S10	SN* Forward	DS was not used	DS was not used
S11	DS	The same search string in the OS	The same dbs as in the OS
S12	DS	The same search string in the OS	The same dbs as in the OS
S13	DS + MS	The same search string as in the OS	The same dbs as in the OS
S14	DS + SN* Back	The same search string as in the OS	The same dbs as in the OS
S15	MS + DS*	A new search string was created	New dbs were used *
S16	DS	The same search string as in the OS	The same dbs as in the OS
S17	MS	DS was not used	DS was not used
S18	DS	The same search string as in the OS	The same dbs as in the OS
S19	OS**: DS + SN Backward	No new search	No new search
S20	DS + SN Forward	The same search string as in the OS	New dbs were added

OS – Original Study; DS – Database Search; MS – Manual Search; SN – Snowballing;

 * – Not used in the OS.

 ** Concerning S20, there was no new search carried out. S20 included 27 additional studies that were not included in its previous review, as they were difficult to locate.

same search strategy as in the previous versions. S9 did not include snowballing despite it being used in the previous version; conversely, S1 and S10 only used snowballing in the updated study. S6 used a subset of the databases used in the previous version, and S4 and S9 used additional databases to the ones used in the previous versions. Finally, S15 changed the search string and S17 changed the search strategy. Further details relating to the search strategy employed by each SLR update are shown in Table 7.

Next, in relation to the quality assessment criteria applied to primary studies, Table 6 shows that twelve SLRs (S3, S4, S5, S6, S9, S10, S11, S12, S13, S14, S15 and S16) used the same criteria as in the original/previous studies. One SLR (S18) adapted its criteria by changing some questions to 'yes/no' questions, and seven SLRs

Table 8
Summary of the SLR method adopted during the updates.

ID	RQ	I/E Criteria	Search Strategy	Quality assessment	Data extraction	TOTAL
S1	■	■	■	–	■	4
S2	□	□	□	–	□	0
S3	□	□	□	□	□	0
S4	■	■	■	□	□	3
S5	□	□	□	□	□	0
S6	□	■	■	□	□	2
S7	■	■	□	–	■	3
S8	□	□	□	–	□	0
S9	□	□	■	□	□	1
S10	□	□	□	□	□	0
S11	□	□	□	□	□	0
S12	□	□	□	□	□	0
S13	□	□	□	□	□	0
S14	□	□	□	□	□	0
S15	□	■	■	□	□	2
S16	□	□	□	□	□	0
S17	■	□	■	–	■	3
S18	□	□	□	■	□	1
S19	□	□	–	–	□	0
S20	■	□	□	–	□	1

Legend: □ Same as in the previous study. ■ Changed in relation to the previous study. - no search/quality assessment

did not apply a quality assessment (S2, S7, S8, S17 and S19 did not apply the quality assessment to both original/previous and current study; S1 and S20 did not apply a quality assessment to the updated SLR). We believe that the decision to not employ any quality assessment to selected studies could have had a detrimental effect upon an SLR update's findings, conclusions and even credibility. Therefore, we believe that the seven SLRs that did not employ any quality assessment have, at least with respect to some aspects, not being well conducted.

Except for S1, S7 and S17, all SLRs adopted the same data extraction form used in the original/previous review. In relation to data synthesis, only four studies (S2, S3, S5 and S11) integrated the results obtained from the inclusion of new studies with the previous SLR results. Fifteen SLRs (S1, S4, S6, S7, S8, S9, S10, S12, S13, S15, S16, S17, S18, S19 and S20) presented the results separately from the previous ones, and S14 included the results only.

Finally, ten SLRs (S2, S4, S6, S7, S9, S12, S14, S15, S16 and S19) involved at least one member of the previous review; seven SLRs (S1, S10, S11, S13, S17, S18 and S20) were performed by a different team, and three SLRs (S3, S5 and S8) were updated by the same team.

Overall, except for the seven SLRs that did not employ any quality assessment of evidence (S1, S2, S7, S8, S17, S19, and S20), all the other 13 SLRs were judged as using valid methods and being well conducted (S3, S4, S5, S6, S9, S10, S11, S12, S13, S14, S15, S16, and S18).

RQ4 - Are there any new relevant methods?

To answer this RQ, the full text of the SLR updates was used to check whether they presented differences from the original versions in the way they searched for evidence, did the synthesis of evidence, whether they included other types of study designs (e.g. qualitative research), and new statistical techniques or improvements to existing techniques used for meta-analysis. This information is available in Tables 6 to 8. A total of nine SLR updates (S1, S4, S6, S7, S9, S15, S17, S18, and S20) have changed some aspect of the method when compared to their previous SLRs' methods. The largest changes were with regard to the search strategies used (six out of nine studies), and research questions employed (five out of nine studies). Conversely, 11 SLR updates did not change their methods, when compared to their previous SLRs (S2, S3, S5, S8, S10, S11, S12, S13, S14, S16, and S19).

RQ5 - Are there any new studies, or new information?

Table 5 shows that the only SLR update that did not carry out new searches on some additional time span was S19. Therefore, all the SLR updates, except for S19, had new studies. We used S19's full-text to check whether there were some new information/conclusions provided (Evidence is provided in Table C.1 in Appendix C); however, none was found.

RQ6 - Will the adoption of new methods change the findings, conclusions or credibility?

This RQ only relates to the nine SLR updates (S1, S4, S6, S7, S9, S15, S17, S18, and S20) that have changed some aspect of the method, when compared to their previous SLRs' methods. We focused upon the evidence provided in Tables 6 to 8 as the basis to answer this RQ. These tables show that S1 and S20 did not apply any quality assessment to selecting new studies. Thus, the credibility of their findings could have been compromised. Furthermore, SLR updates S4, S7, S17, and S20 added a new RQ or sub-question, which can be seen as contributing to increasing the range of findings presented. Therefore, we believe that five SLR updates (S1, S4, S7, S17, and S20) had changes in the findings, conclusions or credibility as the result of changes in the employed methods.

RQ7 - Will the inclusion of new studies/information/data change findings, conclusions or credibility?

To answer this RQ, the full text of both the original SLRs and their updates were used. We provide a detailed table (Table C.1 in Appendix C) reporting the evidence we used from the SLR updates to decide whether findings/conclusions/credibility changed when compared to previous versions. A total of 15 SLR updates (S1, S2, S4, S6, S7, S8, S9, S10, S11, S12, S14, S16, S17, S18, and S20) have changed their findings, conclusions or credibility as the result of the inclusion of new studies/information/data. Only five SLR updates (S3, S5, S13, S15, and S19) did not change their findings, conclusions or credibility as the result of including new studies/information/data.

All the research questions can also be answered using a 'YES'/'NO' associated with each SLR update. Therefore, a summary for the 20 SLR updates, arranged by RQs, is shown in Table 9, below. The table is used later in Section 5 when applying the 3PDF Framework to each of the 20 SLR updates described in this section. An observation is that no single study provided a 'YES' answers to all RQs.

Table 9
Summary of studies according to RQs.

ID	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7
S1	YES	YES	NO	YES	YES	YES	YES
S2	YES	NO	NO	NO	YES	NO	YES
S3	YES	NO	YES	NO	YES	NO	NO
S4	YES	NO	YES	YES	YES	YES	YES
S5	YES	NO	YES	NO	YES	NO	NO
S6	YES	YES	YES	YES	YES	NO	YES
S7	YES	NO	NO	YES	YES	YES	YES
S8	YES	NO	NO	NO	YES	NO	YES
S9	YES	YES	YES	YES	YES	NO	YES
S10	YES	YES	YES	NO	YES	NO	YES
S11	YES	NO	YES	NO	YES	NO	YES
S12	YES	YES	YES	NO	YES	NO	YES
S13	YES	YES	YES	NO	YES	NO	NO
S14	YES	NO	YES	NO	YES	NO	YES
S15	YES	NO	YES	YES	YES	NO	NO
S16	YES	YES	YES	NO	YES	NO	YES
S17	YES	YES	NO	YES	YES	YES	YES
S18	YES	YES	YES	YES	YES	NO	YES
S19	YES	YES	NO	NO	NO	NO	NO
S20	YES	YES	NO	YES	YES	YES	YES

Table 10
Relevance assessment (# respondents) for each of the 3PDF questions.

Step Question	1.a	1.b	1.c	2.a	2.b	3.a	3.b
RELEVANT	8	7	9	9	8	7	6
NOT relevant	1	2	0	0	1	2	3

4. A survey involving authors of slr updates in se

This section describes a survey aimed to obtain feedback from the authors of the SLR updates presented in Section 3 concerning the suitability of using the 3PDF in an SE context.

The objective was to contact the 74 authors (excluding authors of this paper) of the 20 identified SLR updates by email. Unfortunately, four of them were not reachable under the publication e-mail address anymore. Hence, we invited 70 authors to answer the survey, which was instrumented as an online questionnaire implemented in Google forms (see Appendix B). The authors had 15 days to answer the survey, and some had an additional 15-day period. A total of nine responses were received, resulting in an overall response rate of 12.86% (9/70).

Nevertheless, as suggested for survey research (Wagner et al., 2020), we focused on the target population that would be the most appropriate to provide accurate answers, instead of primarily searching for high response rates. Moreover, despite the low number of respondents, the response rate is similar to that documented in other surveys in SE (Nekkanti and Reddy, 2016), and common within the SE field (Ghazi et al., 2018). In addition, the nine respondents were co-authors of eight of the 20 SLR updates, and some provided additional expert feedback on the 3PDF to support our analyses and the discussion on the suitability of using the framework within an SE context.

The questionnaire objectively focused on obtaining feedback on the relevance of each of the 3PDF questions (Steps 1.a-3.c in Fig. 1) and respondents' agreement with the decision drivers (the arrows in Fig. 1). The feedback on the questions and decision drivers is shown in Table 10 and in Table 11, respectively.

It is possible to observe in Table 10 that there was an overall strong agreement on the relevance of the 3PDF questions. Indeed, only three participants (hereafter referred to as A, B, and C) disagreed on the relevance of at least one of the questions. Moreover, only three questions were considered as *NOT relevant* by more than one participant: 1.b, 3.a, and 3.b. We focus the discussion on these.

Regarding question 1.b, which relates to assessing whether an SLR to be updated had good access or use (i.e., if it had an impact on research and/or practice), respondents A and B considered this question as *NOT relevant*. Respondent A mentioned that *"I think it is possible that an SLR that has received a small number of citations and/or has not had a significant impact on practice to become relevant to update"*, while respondent B justified that *"the question might be relevant if considered in an appropriate time frame"*. Nevertheless, we agree with the remaining experts on the relevance of this question as it enables mitigating the risk of investing effort in an SLR update that might not have any impact on research and/or practice. However, we are aware that, as stated by respondent D, *"Relevance to practice might be hard to deduct"*. Also sceptical with regard to question 1.b, Respondent E mentioned that *"While the question is relevant, an SLR might need to be updated even when the answer to 1.b is No, depending on the reasons for the low access or use. E.g., if it is low because of limitations in the original SLR, an update would make sense"*. The 3PDF Framework includes a question that deals specifically with what Respondent E is highlighting – question 2a (Are there any new relevant methods?). A flawed SLR should not be updated (e.g., an SLR that did not use valid methods or was not well conducted), and the 3PDF Framework's question 1.c caters for this. The majority of respondents agree that the 3PDF's question 1.b properly emphasizes that research efforts should be directed towards generating impactful results.

Questions 3.a and 3.b respectively concern the assessment of whether the adoption of new methods or the inclusion of new studies will change the SLR's findings or credibility. Question 3.a was considered as *NOT relevant* by participants A and B, while question 3.b was considered as *NOT relevant* by participants A, B, and C. Respondent A stated that *"The answer to this question may be the outcome of the updated study!"*. Respondent B believes that the findings and credibility would always be affected *"I would always answer these questions YES, no matter what."* Respondent C did not justify her/his answer. In line with respondent A's statement, respondent E, while still considering the questions relevant, mentioned that *"In our case, the update strengthened our results and conclusions. They remain practically the same, so the original ones were valid. This also raised the question of the extent to which the update (and the required effort) was really needed, as the overall picture of the area studied didn't change. Anyway, this can't be ascertained without an update"*. However, as can be seen in Fig. 1, the 3PDF relaxes this condition and requires only one of those questions to be answered with a *YES* or *MAYBE*. Indeed, answering question 3.a requires evaluating the methods employed in the original SLR and comparing them with the intended new methods to use in the SLR update to assess potential effects; while answering question 3.b requires searching for new studies and conducting an informal evaluation of the effects based on reading titles and abstracts. Nevertheless, we agree with the majority of respondents and believe that the upfront effect assessment that should be conducted in steps 3.a and 3.b is beneficial as it avoids investing effort that will not reflect positively on the evolution of the SE evidence corpus.

Table 11 shows the results on the decision drivers. Again, a strong agreement can be observed. Respondents A and B, who were sceptical on some of the 3PDF questions, did also not agree on the decision drivers. Respondent A did not agree with Steps 1 and 3, while respondent B did not agree with Steps 2 and 3. Additionally, respondent E did not agree with the decision driver for Step 1. The reasons are directly related to the previously mentioned statements regarding the questions. Respondent A believes that low access SLRs might become relevant (Step 1) and that instead of an effect assessment, the SLR should be updated (Step 3). Respondent B believes that most SLRs are restricted to a certain time frame (Step 2) and that the findings and credibility would always be affected (Step 3). Respondent E considers that SLRs that

Table 11
Feedback (# respondents) on the agreement with the 3PDF decision drivers.

Decision Driver	ALL questions in Step 1 MUST have a "YES" answer, before moving on to Step 2, when considering ONLY the questions in Step 1 that you judged as RELEVANT?	AT LEAST ONE of the questions in Step 2 MUST have a "YES" answer before moving on to Step 3, when considering ONLY the questions in Step 2 that you judged as RELEVANT?	AT LEAST ONE of the questions in Step 3 MUST have a "YES" answer, when considering ONLY the questions in Step 3 that you judged as RELEVANT?
Agree	**7**	**8**	**7**
Disagree	2	1	2

have low access due to limitations should still be updated (Step 1). Note that methodological limitations (not flawed methodologies) are catered for by the 3PDF's question 2.a. Furthermore, if the use of a flawed methodology has had an effect upon an SLR's access, a totally new SLR should be the solution – one that employs proper methods and research questions. Finally, as previously stated, most respondents agreed that the 3PDF decision drivers are relevant to be employed to decide whether an SLR is suitable for an update.

It is important to note that two of the three respondents who criticized some of the questions in the 3PDF are part of the subset of authors whose SLR updates would not have been assessed as suitable for an update. This issue is also discussed in Section 6 – Threats to validity.

As a result of the survey response analyses, we put forward that the 3PDF was deemed suitable for use within the SE context. Respondent E, for example, argues that *"The framework is relevant and was somehow implicitly used in our decision to update the SLR"*. Respondent F emphasizes that *"Although some aspects of the framework might seem obvious to some people, I think that it can help researchers to decide upon a possible update"*. Indeed, making the 3PDF available to the SE community may help researchers decide on whether to conduct an SLR update and, in case of deciding for an update, to explicitly provide arguments on the rationale employed to make such decision. Moreover, it may serve as guidance for reviewing the relevance of SLR updates. The next section presents an assessment of the suitability for an update of each of the 20 SE SLR updates described in Section 3.

5. Applying the 3PDF to assess the need for updating SLRs

Table 12 shows the results from applying the 3PDF (Fig. 1) to each of the 20 SE SLR updates. This assessment was based upon the summary table already provided in Section 3 – Table 9. Also note that the research questions are directly related to the seven steps part of the 3PDF, as outlined in Section 3:

RQ1 – Step 1.a: Does the published SLR still address a current question?
RQ2 - Step 1.b: Has the SLR had good access or use?
RQ3 - Step 1.c: Has the SLR used valid methods and was well-conducted?
RQ4 - Step 2.a: Are there any new relevant methods?
RQ5 - Step 2.b: Are there any new studies, or new information?
RQ6 - Step 3.a: Will the adoption of new methods change the findings, conclusions or credibility?
RQ7 - Step 3.b: Will the inclusion of new studies/information/data change findings, conclusions or credibility?

The results from applying the 3PDF framework to assess the suitability of an SLR for update showed that only six of the 20 studies would have legitimate grounds for updating a previously conducted SLR (S6, S9, S10, S12, S16, and S18).

Nine original/previous SLRs (See Table 5) had average yearly citations below our cut-off point (S2, S3, S4, S5, S7, S8, S11, S14, and S15), which led to a 'NO' answer to Step 1b. Seven SLR updates did not use any quality criteria to assess studies (S1, S2, S7, S8, S17, S19, and S20), which led to a 'NO' answer for Step 1c. Four SLR updates did not use new methods, which led to a 'NO' answer for Step 2a. Three SLR updates included new methods; however, the inclusion did not change the studies' findings, conclusions or credibility (S6, S9, and S18). This led to a 'NO' answer for Step 3a. Finally, the inclusion of new studies in S13 did not change the findings, conclusions or credibility; this led to a 'NO' answer for Step 3b. Note that the outcome in Table 12 is based on having updates for comparison. It may be more challenging to answer, for example, Step 3b without having done or at least attempted to do an update. Furthermore, the 3PDF framework may inspire researchers to think carefully about ways to add new value with an update, for example, decide to use a new method to motivate an update.

In essence, based on the 20 studies discussed herein, Step 1b ('NO' answer to 9 original/previous SLRs), followed by step 1c ('NO' answer to seven SLR updates), and 3b ('NO' answer to one SLR update) were the steps determining whether an SLR was suitable for an update or not.

Steps 1b and 1c are regarded in the 3PDF as crucial to deciding whether an SLR is suitable for an update, and hence the need for all answers to be 'YES' to proceed to Step 2. These steps can be summarised as: Are the SLR findings contributing to research and practice? Can we trust the results that are to be provided in the SLR update?

Note that Step 1b is driven by how relevant the original/previous SLR has been to the research community, or to practice, or to both, based on the average yearly citation and/or downloads. The decision as to which topic to select as an SLR can be informed, for example, by a preliminary mapping study on the topic, so to find out existing gaps and also whether such mapping study has had good access or use.

Note that, as shown in Table 5, six of the nine SLR updates associated with the nine original/previous SLRs that presented an average yearly citation below our cut-off point, were published up to only one year after the publication of the original/previous SLRs. Such a short time span may explain the low average yearly citations shown for the original/previous SLRs.

With regard to Step 1c, methodological rigor is expected and should be employed in every SLR, regardless of it is an update of a previously published SLR.

With regard to Step 3b, it addresses the question as to whether the new research findings in the SLR update would make any significant contributions to the state-of-the-art when compared to what was already known via the original/previous SLR. The answer to this step requires some effort as one needs to carry out a few preliminary searches to perform a risk analysis on whether the SLR update's research contribution motivate an update. However, this is still worth doing in order to at least reduce the number of SLR updates that are being published that are not needed, according to the 3PDF framework.

Table 12
Application of the updated 3PDF to previous SLR updates in SE.

SID	1.a	1.b	1.c	2.a	2.b	3.a	3.b	Recommendation
S1	YES	YES	NO	–	–	–	–	No: Update not needed
S2	YES	NO	NO	–	–	–	–	No: Update not needed
S3	YES	NO	YES	–	–	–	–	No: Update not needed
S4	YES	NO	YES	–	–	–	–	No: Update not needed
S5	YES	NO	YES	–	–	–	–	No: Update not needed
S6	YES	YES	YES	YES	YES	NO	YES	Yes: Prepare update
S7	YES	NO	NO	–	–	–	–	No: Update not needed
S8	YES	NO	NO	–	–	–	–	No: Update not needed
S9	YES	YES	YES	YES	YES	NO	YES	Yes: Prepare update
S10	YES	YES	YES	NO	YES	–	YES	Yes: Prepare update
S11	YES	NO	YES	–	–	–	–	No: Update not needed
S12	YES	YES	YES	NO	YES	–	YES	Yes: Prepare update
S13	YES	YES	YES	NO	YES	–	NO	No: SLR up to date
S14	YES	NO	YES	–	–	–	–	No: Update not needed
S15	YES	NO	YES	–	–	–	–	No: Update not needed
S16	YES	YES	YES	NO	YES	–	YES	Yes: Prepare update
S17	YES	YES	NO	–	–	–	–	No: Update not needed
S18	YES	YES	YES	YES	YES	NO	YES	Yes: Prepare update
S19	YES	YES	NO	–	–	–	–	No: Update not needed
S20	YES	YES	NO	–	–	–	–	No: Update not needed

Step 1.a Does the published SLR still address a current question?
Step 1.b Has the SLR had good access or use?
Step 1.c Has the SLR used valid methods and was well-conducted?
Step 2.a Are there any new relevant methods?
Step 2.b Are there any new studies, or new information?
Step 3.a Will the adoption of new methods change the findings, conclusions or credibility?
Step 3.b Will the inclusion of new studies/information/data change findings, conclusions or credibility?

6. Threats to validity

6.1. Threats related to identifying the se slr updates and assessment

The four types of validity threats and corresponding aspects of threats that are discussed in this sub-section (see Table 13) have been identified by Zhou et al. (Zhou et al., 2016) as the most discussed threats and aspects in SE SLRs, based upon data extracted from 316 SE SLRs published within the period 2004 to 2015.

Four types of validity threats have been identified (Zhou et al., 2016):

Internal (In): Unknown factors that may affect the dependent variable. E.g., confounding factors we are unaware of.
External (Et): To what extent we can generalize the findings.
Conclusion (Cn): To be able to draw correct conclusions regarding the relationships being investigated, e.g. use of an adequate statistical test, use of proper measurement
Construct (Ct): Represents to what extent operational measures precisely measure the concepts they claim to measure.

The aspects of validity threats that are discussed herein are the following (Zhou et al., 2016):
Incorrect or incomplete search terms employed in database searches: The search string may include inadequate search terms related to the SLR topic, and/or not include terms that should have been included. Associated with Construct and Internal validity.
Incomprehensive venues or databases: Some important databases are not available for not being subscribed by the Institutions to which the authors are affiliated. Associated with Construct validity.
Bias in study selection: While searching for candidate studies, researchers select studies subjectively, without completely complying with the inclusion and exclusion criteria to support their decision. Associated with Internal and Conclusion validity.
Misclassification of primary study: "Researchers may not consider all the classifications at the beginning of the research process or the primary studies selected by the reviewers are mistakenly classified" (Zhou et al., 2016). Associated with Internal and Conclusion validity.
Publication bias: Bias associated with the problem that positive results are more likely to be published than negative results (Kitchenham, 2004). Associated with Internal validity.
Bias in data extraction: While extracting data, researchers may not totally understand what needs to be extracted, and how the data relates to the research questions. Associated with Internal and Conclusion validity.

6.2. Threats related to the survey with se slr update authors

We identified some threats throughout the process of planning and conducting the survey. Table 14 lists these potential threats, classified by the survey validity types presented in (Linåker et al., 2015), and explains how they were addressed in our survey. One of the main mitigation strategies was validating the instrument amongst the authors of this paper, before handing it over to the SE SLR update authors. All four authors of this paper have carried our numerous SLRs and have also participated in SLR updates.

There was an additional threat that relates to a possible bias in the respondents' answers to motivate the SLR update(s) they co-authored. Out of the nine respondents, four co-authored SLR updates that would have been updated if the 3PDF framework had been used; all of them were very supportive of the Framework's questions. Conversely, the only three respondents who were critical of the Framework would not have had their SLR updates selected if applying the 3PDF framework. However, another two respondents, who would also not have had their SLR updates selected if applying the 3PDF framework, were very supportive of the Framework's questions. Therefore, we argue that the evidence presented here cannot be argued to suggest a bias for or against the use of the 3PDF framework.

Table 13
Validity & Aspects of Threats and Treatments.

Aspects	Treatment
Incorrect or incomplete search terms employed in database searches	We carried out a manual search in the proceedings of two conferences – ESEM (International Symposium on Empirical Software Engineering and Measurement) and EASE (International Conference on Evaluation and Assessment in Software Engineering), for the period 2006 to 2018, based upon titles, keywords and abstracts. These conferences were selected for having a clear focus on Evidence-based Software Engineering, and for including secondary studies as some of the topics of interest. A total of 909 papers were screened. Later, we also added synonyms to search terms, prior to carrying out the database searches.
Incomprehensive venues or databases	First, the databases selected for use with the database searches were: ACM Digital Library; IEEE Xplore; ScienceDirect; Scopus and Springer Link. These databases were selected for being databases recommended by previous SLRs as the most adequate online databases to use (e.g. Maplesden et al. (Maplesden et al., 2015), and Zhang et al. (Zhang et al., 2011)). Note that there are other databases (e.g. EI compendex, Wiley Interscience (Wiley Online), Inspec and Kluwer) that were not selected due to the high degree of overlap between the results and those for the selected databases (Maplesden et al., 2015).
	Second, we had access to all the above mentioned databases through our Institutions' subscriptions.
Bias in study selection	
	To identify terms for the search string, the third author carried out a manual search on proceedings and later discussed the selection with the first author, and a full agreement was reached.
	Both the first and third authors searched independently through reading titles and abstracts for the candidate studies retrieved via database searches and later compared their respective selection until consensus was reached. Later, both the first and third authors looked at the full text of the 27 studies that were selected from the first phase and selected 20 with a high level of agreement.
	Finally, both authors have experience carrying out a number of SLRs.
Misclassification of primary study	With regard to this work, the classifications corresponded to the different steps part of the 3PDF framework, and also represented by this SLR's research questions (RQs). The data needed to answer the RQs was extracted by the third author and always discussed with the first author until consensus was reached. Thus, we had two researchers working through each of the SLR updates, and we do believe that such process mitigated classification biases.
Publication bias	Within the context of this SLR, any SE SLR updates identified via our searches were included, regardless of whether they included additional evidence from more recently published primary studies and/or new conclusions.
Bias in data extraction	Bias could have been introduced given two co-authors in this paper (first and fourth) also co-authored one of the SLR updates.
	The second issue was addressed by forming a studies selection team that included one author (third author) who did not co-author any SLR update, and who was also responsible for extracting the data from the 13 SLR updates that were selected.

Table 14
Threats to Validity and Treatments.

Threats	Treatment
Face Validity – Bad instrumentation	Revision and evaluation of the questionnaire about the format and formulation of the questions. Questions objectively focused on the 3PDF. Running a pilot study.
Content Validity – Inadequate explanation of the constructs	Revision and evaluation of the questionnaire about the format and formulation of the questions. Running a pilot study. Providing a brief explanation on the 3PDF and a link with further details.
Criterion Validity – Not surveying the target population.	We identified SE SLR update authors following an explicitly documented and carefully conducted procedure (cf. Section 3).
Construct Validity – Inadequate measurement procedures and unreliable results.	We only used frequency counting, which can be safely applied to discrete survey questions concerning the relevance of the 3PDF questions and the agreement with the 3PDF decision drivers. Also, we triangulated the answers with the provided explanations.
Reliability – Lack of statistical conclusion validity	This threat strongly depends on the sample size. Unfortunately, while contacting twice the SE SLR update authors we were aware of, our final sample size was still limited. Hence, we focused our results on qualitative analyses and did not make any further claims on conclusion validity.

6.3. Threats related to applying the 3PDF to assessing the se slr updates

In relation to validity threats with regard to applying the 3PDF to assessing the SE SLR updates, we believe that the three main validity issues herein relate to whether: i) authors misunderstood the 3PDF framework; ii) bias being introduced given two co-authors of this paper (first and fourth) also co-authored one of the SLR up-

dates; iii) bias being introduced when applying the 3PDF framework to the SLR updates.

With regard to the first issue, the original paper detailing the 3PDF Framework was shared and discussed between all the authors, to solve any doubts/misunderstandings/etc. The 3PDF Framework was also used as a basis for the instrument and corresponding survey, which are detailed in Section 4. None of the survey's respondents mentioned anything related to ambiguity in regard to

Table 15
Summary of the SLR methods adopted during the updates.

ID	RQ	I/E Criteria	Search Strategy	Quality assessment	Data extraction	TOTAL
S1	■	■	■	–	■	4
S2	□	□	□	–	□	0
S3	□	□	□	□	□	0
S4	■	■	■	□	□	3
S5	□	□	□	□	□	0
S6	□	■	■	□	□	2
S7	■	■	□	–	■	3
S8	□	□	□	–	□	0
S9	□	□	■	□	□	1
S10	□	□	□	□	□	0
S11	□	□	□	□	□	0
S12	□	□	□	□	□	0
S13	□	□	□	□	□	0
S14	□	□	□	□	□	0
S15	□	■	■	□	□	2
S16	□	□	□	□	□	0
S17	■	□	■	–	■	3
S18	□	□	□	■	□	1
S19	□	□	–	–	□	0
S20	■	□	□	–	□	1

Legend: □ Same as in the previous study. ■ Changed in relation to the previous study. - no search/quality assessment

the survey's instrument, which we believe can also be used as additional evidence showing that the 3PDF Framework was understood by the authors.

In relation to the second issue, there was only one SLR update that has been co-authored by the first and fourth co-authors in this paper (S9). All the summary Tables 5 to 8, and also C.1 in Appendix C were prepared by the third author, and later partially re-checked by the first author (e.g. text in Table C.1, all of Table 8, naming consistency in Table 5, and adding average yearly citation). The content was used to answer the research questions outlined in Section 3, where we also provide details on how each RQ has been addressed. Further, Table 9 provides a summary for all the 20 studies, and the corresponding answers 'YES' or 'NO' can be checked against the summary data shown in Tables 5 to 8, Table C.1, and the explanations provided about how each RQ was answered. Section 3 provides the data needed to assess each of the 20 SLR updates. In fact, Table 9 was directly used in Section 5. Finally, study S9 was assessed by the third author.

Finally, as for the third issue, Section 3 provides research questions that match all the steps part of the 3PDF Framework directly, and Table 9 can be used as input to the assessment shown in Table 12. Thus, we believe that a good quality Section 3 is directly related to how well the assessment was done in Section 5.

Section 3 provides summary tables (Tables 5 to 8, and Table C.1 in Appendix C), and details how the tables were used to answer each RQ. Finally, an overall summary table (Table 9) is also provided, so to enable anyone to check it against the summary data shown in Tables 5 to 8, Table C.1, and the explanations provided under each RQ on how the RQs were answered. These tables were prepared by the third author based on the full-text for each of the 20 studies and were later on partially re-checked by the first author (who also used the full-text for the included studies). Once the assessment was done, basically via using Table 9 to prepare Table 12, the second and fourth authors evaluated the results to see if anything unusual/mistake could be identified. None were found.

7. Implications

The results from applying the 3PDF Framework to assess the suitability for updating the 20 SE SLR updates showed that half of the 14 SLR updates assessed as unnecessary to update were published within a time span of no more than two years from the publication of the original/previous SLR. These are: S2 – mutation testing, S3 – cloud computing, S4 – safety certification, S5 – software process improvement, S14 – open-source development, S15 – process simulation, and S19 - requirements elicitation techniques. None of the SLRs assessed as suitable for updating were published within a time span of no more than two years from the publication of the original/previous SLR. The shortest time span was three years, for SLR updates: S12 (maintainability prediction), S13 (motivation in SE), and S16 (software ecosystem). We believe that one of the several take away messages for the SE research community is that the 'maturation' time between the publication of the original/previous SLR and its update should be longer than two years, although exceptions exist. Such suggestion does not however bypass the need to apply the 3PDF to assess whether an SLR update is genuinely needed. Another take away message is that our results show that many resources are wasted in carrying out updates that did not need to be conducted. This applies to all the 14 SLR updates that were assessed as unsuitable for an update, not only the eight abovementioned.

Furthermore, the results detailed in Section 5 also illustrate that several SLR updates did not employ any quality control when selecting primary studies, which has clear detrimental implications to the quality of the evidence being provided by these SLR updates. This applied to seven SLR updates (S1 - 2017, S2 - 2019, S7 - 2018, S8 - 2015, S17 - 2014, S19- 2008, and S20 - 2017). It is a surprising result given that there are numerous guidelines and literature on how to carry out SLRs in SE, and they all stress the use of quality checklists (e.g. (Kitchenham, 2004, Kitchenham and Charters, 2007)). Note that, except for S19, all the other six SLR updates were published in the past five years, thus quite recently. As pointed out in Section 1, there are now hundreds of SE SLRs published. Therefore, another take away message is that we believe that it is important that not only SLR authors, but also reviewers of SLRs submitted to conferences and journals should be familiar with SLR guidelines to avoid the publication of SLRs that are methodologically flawed.

Another important aspect that has been brought to light is that, whenever SLR updates do not include new relevant methods, they

should at least provide results that reflect a change in findings, conclusions, or credibility, when compared to the original/previous SLR. The point here is that the effort to carry out an SLR update should be justified by some novelty, in either the methods used or new findings/conclusions/credibility. If nothing has changed, one may as well rely on the findings already presented in the original/previous SLR. This would also apply to newly starting PhD students, i.e., they should not waste any effort to carry out an SLR update if such update is not justified by a novelty in either the methods used or new findings/conclusions/credibility. And the justification for the decision to not update (or to update) an existing SLR can be based upon applying the 3PDF to the original SLR, and additionally supported by the evidence and recommendations provided herein.

This paper has focused on the "When" to update SLRs. However, we also intend to tentatively address the "How" to update SLRs. We have reproduced here Table 8 (shown as Table 15 below). It highlights the six SLR updates that we assessed as suitable for an update, as per the 3PDF framework. Four of these used exactly the same RQs, I/E Criteria, search strategy, quality assessment and data extraction as used in the original/previous SLRs (S10, S12, S16, and S18). S9 used additional databases for the database searches, and S6 used a subset of the databases. Finally, S6 added a new I/E criterion based on publication date. Based on the evidence, we believe that an additional take away message would be to employ the same mechanisms used in the original/previous SLR being updated. This was the case in four of the six SLR updates discussed herein. In addition, we have made recommendations elsewhere as to what search strategies to employ when updating SE SLRs (Mendes et al., 2019), using one of the SLR updates presented herein (S9), and two of its replications. These can complement the suggestions abovementioned regarding "How" to update SLRs in SE.

There is one last comment we would like to make. The research process that we have employed herein could also be seen as a set of Design Science (DS) activities: i) problem analysis, which is characterised by our assessment relating to the suitability for update of SE SLR updates; ii) solution design, which is characterised by using the 3PDF framework; and iii) solution evaluation, which is characterised by our survey. We did not carry out the research detailed herein with DS in mind. However, we believe that it is important to stress that DS would also have been a suitable match for our research process.

8. Conclusions

This paper investigates and makes recommendations towards an important aspect relating to the update of SLRs in SE – when to update an SLR. We initially formulated four overall research questions in Section 1. To summarize, ORQ1 is addressed by presenting some proposals for how to decide when to update an SLR in Section 2. Next, ORQ2 is addressed by conducting a systematic search for updates in SE, which is presented in Section 3. ORQ3 is addressed through a survey of authors having published updated SLRs in SE. The survey is presented in Section 4. Finally, ORQ4 is addressed by assessing the retrieved updates of SLRs using the 3PDF framework. The outcome is presented in Section 5.

In more detail we address the challenge of deciding when to update an SLR using three consecutive stages: Stage 1, explained in Section 3, details a two-step process (see Fig. 3) used to identify relevant SLR updates in SE, which provide not only the set of 20 SE SLR updates to be employed in Section 5, but also the authors, who were contacted via a survey and invited to provide their feedback about the suitability of a third-party decision framework (3PDF), proposed in (Garner et al., 2016), to be used by the SE community to assess the suitability of SE SLRs for updating. The survey – our Stage 2, is detailed in Section 4. Nine authors answered the survey,

out of the 74 who were contacted twice by email. As Stage 3, we have applied the 3PDF framework to assess the suitability of each of the 20 SE SLR updates for updating (detailed in Section 5).

After applying the 3PDF, results showed that 14 of the 20 SE SLR updates did not need updating. Based on the assessment of the 20 SLR updates, the main decision driver to whether an SLR should be updated or not was Step 1b - the SLR's contribution to research and practice, which was given a 'NO' answer to nine of SLR updates. Another two steps – Steps 1c and 3b, received 'NO' answers to another seven and one SLR updates, respectively.

The results, in our view, suggest that the use of a decision support mechanism (as the 3PDF) to help the SE community decide upon the need for SLR updates is very important. We also put forward that the 3PDF should be adopted by the SE community to keep relevant evidence up to date and to avoid wasting effort with unnecessary updates.

One of the avenues of future work is to also contact authors of SE SLRs, which have not been updated yet, in order to obtain their assessment on the suitability of the 3PDF for use in SE.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Seven excluded studies listed in Table 4.

S21 A. B. Soomro, N. Salleh and A. Nordin "How personality traits are interrelated with team climate and team performance in software engineering? A preliminary study" 9th Malaysian Software Engineering Conference (MySEC' 15), p. 259–265, 2015.

S22 O. Matos, T. Conte and E. Mendes, "Is there a place for qualitative studies when identifying effort predictors?: a case in web effort estimation", 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14), Article 40, p. 1–10, 2014.

S23 O. Dieste, T. Marta and L. F. Ramos, "Obtaining Well-Founded Practices about Elicitation Techniques by Means of an Update of a Previous Systematic Review", 20th International Conference on Software Engineering and Knowledge Engineering (SEKE' 08), p. 769–772, 2008.

S24 J. A. Oliveira, M. Vargas and R. Rodrigues, "SOA Reuse: Systematic Literature Review Updating and Research Directions", 14th Brazilian Symposium on Information Systems (SBSI' 18), Article 71, p. 1–8, 2018.

S25 C. Gao, S. Jiang and G. Rong, "Software process simulation modeling: preliminary results from an updated systematic review", 2014 International Conference on Software and System Process (ICSSP' 14), p. 50–54, 2014.

S26 H. Zhang, B. A. Kitchenham and D. Pfahl, "Software Process Simulation Modeling: An Extended Systematic Review", New Modeling Concepts for Today's Software Processes (ICSP' 10), Lecture Notes in Computer Science, Volume 6195, 2010.

S27 R. Bezerra, F. da Silva, A. Santana, C. de Magalhães and R. Santos, "Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study", Intl. Symp. on Emp. Softw. Eng. and Measurement., pp. 1–4, 2015

Appendix B

The survey questionnaire contains an introduction and three steps to provide feedback on each of the decision framework questions and decision drivers. The survey questions provided in each of the three steps, as presented to the respondents, are shown in Fig. B.1.

Fig. B.1. Survey steps and questions as presented to the respondents.

Appendix C

This Table provides evidence gathered from each of the SLR updates and used to answer step 3b in the 3PDF (Table C.1).

Table C.1
Changes in SLR's conclusions? (Step 3B).

ID	Publication timespan	Changes in conclusions?	Evidence
S1	7 years	Yes	"Results: … Eleven [M2, M3, M4, M6, M7, M10, M12, M13, M14, M15, M16] out of the 17 pa-pers that we have examined propose new models…" "Results: .. Among the 11 new models, we have identified a first group [M2, M3, M12] whose main concern is to address SRP in agile contexts …" "Results: A second group of new models seems more concerned in proposing … For the remaining new models, …" "Results: In the case of the new SRP models, we report three main categories of outputs …" "Discussions: In contrast with the results reported in (Hoisl and Sobernig, 2016), we have found more proposals aimed to scale in presence of large sets of requirements, …" "Discussions: "It is interesting to analyze the evolution of the SRP field by comparing the results from (Hoisl and Sobernig, 2016) and ours … In our literature review, we have found 6 papers …"

(continued on next page)

Table C.1 (*continued*)

ID	Publication timespan	Changes in conclusions?	Evidence
S2	1 year	Yes	"Discussion: Svahnberg et al. (Hoisl and Sobernig, 2016) found that most models focused on a limited set of input fac-tors, mainly hard constraints. Only a 57,1% of the reviewed models considered soft factors. In our study, 15 out of the 17 models (88,2%) do include soft factors …" "Conclusion: … The main results (detailed in Section 4) show some progress with respect the previous proposals in the period 1997–2008 surveyed in (Hoisl and Sobernig, 2016), in particular: …" "Abstract: Researchers can use our results to find more detailed information about particular techniques, and to design comparable and reproducible experiments. …." "Introduction: We previously presented preliminary results from our SLR in a workshop paper (Ferrari et al., 2018a), with general classifications of primary studies. This paper greatly extends and updates the prior short paper, and makes four key contributions: • …" Section 3.3. "Step 4–Post-publication update: Our set of selected studies was updated after our preliminary publication (O. Dieste et al., 2008)." "Section 6: This section presents more details about the measurements the studies used …" "Conclusions: We updated and extended our prior paper (Ferrari et al., 2018a), analyzed the evolution of research on the topic, and summarized its underlying goals, techniques, metrics used, and results achieved …" "Conclusions: Based on the research questions investigated in this work, we summarize several recommendations. …" "Conclusions: We also found several studies that combined techniques to achieve substantial savings. For example, …"
S3	0 year	No	Original study: "Conclusion – … a significant part of the companies perceived inclination towards for the innovation adoption process influenced by technological, organizational and environmental contexts … it is expected that the results in this systematic literature review would help in establishing knowledge on how the companies should adopt and migrate to the cloud, how the cost-benefit relationship can be evaluated as well as providers can be selected. These findings can be a useful reference to develop guidelines for an effective use of cloud computing …" Update study: "Conclusion – … a significant part of the companies perceived inclination towards the innovative adoption process influenced by technological, organizational and environmental contexts … The results in this systematic literature review can help the development of guidelines to support newcomers companies to adopt and migrate to the cloud, how the cost-benefit relationship can be evaluated as well as the selection of providers …"
S4	1 year	Yes	"Abstract: … As a major finding of the review, the results strongly suggest the need for more practitioner-oriented and industry-driven empirical studies in the area of safety certification…." "Introduction: .. we classify into a hierarchical taxonomy … The taxonomy includes 49 basic evidence types and is, to our knowledge, the most comprehensive classification of safety evidence built to date." "Introduction: … Our study notably indicates that a large majority of the approaches surveyed have not been validated in realistic settings and thus provide little information about their practical utility …" "Introduction: … The work we present here extends an earlier conference paper (Maplesden et al., 2015). The main extensions are: (1) the addition of a new data source, namely Google Scholar, (2) significant expansion of the description of the research method and the results; and (3) our domain analysis …" "Related work: … There are a number of SLRs in the literature whose scope partially overlaps with ours, … None of these specifically address the topic of evidence for safety." "Results: With respect to the application domains and the safety standards referred to in the studies, we identified eight application domains and 16 safety standards." "4.1 … RQ1: We created a taxonomy, shown in Fig. 2, for evidence types based on the various evidence examples, artefacts, tools and techniques found in the primary studies …" "4.1 … RQ1: Our results indicate that the most frequent evidence types referred to in the literature are Hazards Cause Specification …" "4.2 … RQ2: … we identified some technique for structuring safety evidence. We divide the techniques into three main categories …" "4.3 … RQ3: … We identified techniques for evidence assessment … We classify these techniques into four categories …" "4.4 … RQ4: … We identified several categories of general challenges and needs related to providing safety evidence information and to structuring and assessing the evidence …" "Discussion: … The results from the review provide a general research-oriented view on evidence provision … We believe that this taxonomy is a useful reference to new researchers, helping them get better acquainted with the area" "Discussion: … Alongside the taxonomy, our results concerning evidence structuring and assessment serve as useful input for future work on tool support …" "Discussion: … We believe that more research on safety evidence at lower levels of abstraction (system type level and specific system level) is necessary …" "Discussion: … The results regarding evidence structuring (RQ2) are useful for both research and practice to promote further work on managing large collections of evidence data …" "Conclusions and Future Work: … One of the main outcomes of the SLR is a general taxonomy of safety evidence type …" "Conclusions and Future Work: … The SLR further examined and classified existing techniques for structuring evidence information into three categories …" "Conclusions and Future Work: … the paper presented a comparison of eight safety-critical domains …" "Conclusions and Future Work: … The SLR provides useful insights for both researchers and practitioners…"
S5	2 years	No	"Abstract: The results of this review, like our previously conducted systematic review (Sulayman and Mendes, 2009), show that very few studies have specifically focused on SPI …" "Research Gaps: … Similar to our previous execution …, the number of studies that met our inclusion criteria is small …"

Table C.1 (*continued*)

ID	Publication timespan	Changes in conclusions?	Evidence
			"Research Gaps: ... Studies S1, S4, S5, S6, S7 and S8 propose generic SPI models ... However, none of the studies has suggested a specific model for Web SPI ..."
			"Research Gaps: ...The systematic review did not identify any SPI model or technique specifically customized for Web companies ... However, this should not be understood as indicating that there is no need for specific Web SPI models and techniques. In fact, previous work on Web engineering may suggest the contrary."
			"Research Gaps: ... We argue that in terms of SPI it may also be the case that research on result oriented, cheaper and less time consuming SPI strategies for small and medium Web companies need to be conducted ..."
			"Conclusion: Due to our strict inclusion criteria, the number of relevant studies found was small but the overall search process was very comprehensive ..."
			"Conclusion: Found studies suggested did not suggest any specific model or technique made to measure for the SPI of Web companies ..."
			"Conclusion: The review as mentioned earlier helped us in identifying the possible research gaps and directions."
S6	4 years	Yes	"Introduction: ... Like the previous review, the current review recognises the heterogeneity of the research in this relatively new field and aims to provide some organization and structure to this area."
			"Results: 18 papers described online games both for entertainment (9) and learning (9), suggesting that these have become more popular since the previous review in 2012, when only 9 were found."
			"Results: 21 studies in the current review provided additional evidence to support previous research documenting the attentional ..."
			"Results: ... the current update to the systematic review of Connolly et al. found many more papers reporting empirical evidence of the positive outcomes of playing games (512) than the previous review (129). This illustrates the increased interest in the positive impacts of digital games during the five year period from 2009 to 2014 compared to the previous five year period."
			"Discussion: More high quality papers using the term "serious games" were identified in the current review (34) than in the previous review (8) suggesting that the term has entered the mainstream in recent years ..."
			"Discussion: As in the previous review, similar numbers of papers about entertainment games and games for learning were found, but these studies differed in their aims, the platform used for delivery and the learning and behavioral outcomes they addressed."
			"Discussion: ... Some progress is being made in understanding how specific game features engage players and support learning."
			"Discussion: Progress has been made too in understanding features that make games more engaging ..."
S7	5 years	Yes	"Introduction: This study contributes to the cloud evaluation literature by: - providing an updated review and synthesis of the CSEM literature between 2006 and 2016 identifying research deficiencies of CSEMs for suitably positioning new research activities ..."
			"Related Work: We have improved our preliminary SLR by (1) extending the publication period and defining an improved search string to provide an increased number of retrieved related papers, (2) applying a new data extraction dimension to provide more useful information regarding the CSEM components, and (3) examining the CSEMs based on MCDM making the review more focused. This focus will facilitate building even closer ties between the cloud evaluation context and MCDM context by identifying and understanding the associated MCDM method problems and deficiencies."
			"Conclusion: Presenting new lines of inquiry that could inspire researchers to conduct more research within the area of cloud service evaluation is a main contribution of this SLR ..."
			"Conclusion: These research deficiencies in CSEMs emphasize the requirement to rethink the proposed methods, components, and processes in evaluating cloud services."
S8	3 years	Yes	"Abstract: ... we discovered 31 different CSFs for ERP system implementation projects."
			"Motivation: ... several case studies, surveys, and literature reviews have already been conducted by different researchers (e.g., (Kitchenham et al., 2004, Hoisl and Sobernig, 2016, Gao et al., 2014)). Most of these literature reviews cannot be reproduced, because descriptions of the review methods and procedures are lacking ..."
			"Critical Success ...: The goal of the performed literature review is to gain an in-depth understanding of the different CSFs already identified by other researchers."
			"Literature Review – Results: In most previous literature reviews, the CSFs were grouped more coarsely so that a lower number of CSFs was used ... we used a larger number than earlier researchers had because we expected the resulting distribution to be more insightful."
			"Literature Review – Results: Comparing these results with other literature reviews, ..., it became obvious that the top five factors were similar while only the ranks differed. Due to our large literature base, our total numbers are much higher ..."
			"Literature Review – Results: ... This clearly is a research gap in the ERP CSF research area ..."
			"Conclusion: Our study reveals that there are several papers – case studies, surveys as well as literature reviews – that are focusing on CSFs."
			"Conclusion: ... we derived 31 different CSFs. We identified the following top three CSFs ..."
			"Conclusion: This ranking is similar to the ranking of other literature reviews (e.g., (de Paula and de Carneiro, 2016, Hoisl and Sobernig, 2016)). Compared to these other reviews, the number of papers included in our study exceeds their numbers. ..."
S9	7 years	Yes	"Partial Results: This difference relates to an increase in the number of cross-company models to at least two ... This was not observed previously, when the original SLR was carried out."
			"Discussion: ... in addition to the within-company dataset sizes may have influenced the changes in the patterns that were previously documented in the original SLR on which this research is based upon ..."
			"Discussion: ... Contrary to the results detailed in (Kitchenham et al., 2004), our results provide some support to that supposition, thus suggesting that time spent in quality control ..."
			"Conclusion and Future Work ...We believe that this newly observed behavior might be related to the ..."

Table C.1 (*continued*)

ID	Publication timespan	Changes in conclusions?	Evidence
S10	4 years	Yes	"Abstract: We identified a strong indication of solutions based on Artificial Intelligence … We also identified that there is a gap in terms of agreement on suitable cost drivers … " "Abstract: This updated review of the state of the art resulted in 24 new relevant papers selected …" "Introduction: … we applied the Forward Snowballing technique (Kitchenham et al., 2006) to ¯nd out relevant studies since the reference review of Usman et al. (de Paula and de Carneiro, 2016)." "Introduction: As contribution of our review, 24 new relevant papers were selected. Some findings from the reference review remain actual, but other questions have been raised in our research …" "Conclusion: Some considerations of the reference SLR are still valid and current. However, we identified new trends."
S11	3 years	Yes	"Abstract: The main contribution of this study relates to the identification of other practices, including sharing of the "social conscience", …" "Introduction: This study extended the studies on this subject, considering from 2012 to 2014 year and also including 11 new studies found in current literature…" "Results: Collaborative Technology practice was the most mentioned in both SLR. So, it was selected as an important practice." "Results: Another limitation cited by the Arshad et al. (2012) is the … This finding was supported in the study case presented in Zahedi and Babar (2014), where an attempt …" "Results: Another practice detected in the selected works is Knowledge Sharing, … This practice is also mentioned in the KM literature as a way for knowledge conversion known as "socialization" …" "Conclusion: The study identified a 35% increase in volume related to KM practices in GSD if compared to the primary study …" "Conclusion: It was also observed that the practices "common understanding" and "knowledge sharing" are even more frequent in the work of the past 12 years…" "Conclusion: … Among the contributions of this work are the identification of KMP Shared social context, and the remark about the need and the importance to use new approaches to improve informal communication …"
S12	3 years	Yes	"Discussion: An interesting finding of the SR is that no predictors of maintainability or factors that impact upon maintainability were found that …" "Discussion: The comparison of the two SRs suggests that findings differ in terms of use of maintainability prediction techniques, importance of the predictors, and the project stage when the predictors are gathered." "Discussion: Therefore, one can argue that the comparison given in Section IV would have yielded different results if the previously conducted SR was extended to include recently published studies."
S13	3 years	No	"Abstract: … There is no significant advance in the last 5 years …" "RQ1: … The Original Study reached similar results…" "RQ2: … The only new de-motivator that appeared in this study was …" "RQ3: … Only one external sign was added to the original list …" "RQ4: … All aspects identified in the current study are described in the original study, but two: … are cited in only one paper each …" "RQ5: … The MOCC model … constructed based on the findings from the original study, has been cited twice." "Discussion: The findings of our review confirm the results from the original study …" "Discussion: … several motivators in V did not appear as motivators in the original study." "Discussion: Regarding the motivating aspects of software engineering, the current study agrees with the original study, and adds … as motivating aspects …" "Discussion: (JCT)' is the most recurrent model in the original study, but it was found only in two papers in the current study…" "Discussion: Our results suggest that the overall understanding of how software engineers are actually motivated does not seem to have significantly advanced in the last five years …" "Discussion: In both the current and the original studies, survey research represents the majority of empirical primary studies." "Limitations of this review: The problem of cultural bias faced in the Original Study is also present in this study …" "Results: This study has identified new descriptors for software engineers' characteristics, and new motivators and demotivators for software engineers, besides those described in the Original Study."
S14	0 year	Yes	"Abstract: In this paper, we extend a systematic literature review (SLR) to cover decisions with respect to model driven development (MDD) tooling. We extract tooling information from the selected SLR papers and classify these software tools supporting DSML development …" "Section 3: In this section, we report on the results of the SLR regarding DSML tooling support. …" "Section 3.2: In previous work [12, (Hoisl and Sobernig, 2016)], for each DSML design found documented in the publications of our paper corpus, we identified and recorded the decision options for six decision points (see Section 2.2). …" "Section 3.2: From the above sources, we extracted tooling information for each of the eight categories (see Section 3.1) …" "Section 3.2: In Table 3, we differentiate between open- and closed-source license models per tooling category …" "Conclusion: In this paper, we presented an extension to a SLR study to include an additional decision point of tooling support for UML-based DSML development …" "Conclusion: These empirical findings complement recent, qualitative research on MDD tool adoption (Kitchenham et al., 2007)."
S15	1 year	No	"Abstract: The review identified 42 primary studies from 1992 to 2013. … The overall findings confirmed the positive impact of SPS on education" "Introduction: To achieve the goal, we extended our previous Systematic Literature Review (SLR) (Gao et al., 2014, Hoisl and Sobernig, 2016, Ghazi et al., 2018) on this more specific topic …" "Search Process: Our previous work (Ghazi et al., 2018) covered non-game simulation to a large extent, … Hence, there was a need to carry out a search using the identified important games."

Table C.1 (*continued*)

ID	Publication timespan	Changes in conclusions?	Evidence
			"Related Work: In summary, our work differs from the previous related endeavors in a number of aspects. In particular, we set a different review scope from others with special focus on process simulation and explicitly include non-game simulators; we paid special attention on how to support SE educators to successfully adopt SPS in their curricula, which was rarely raised before."
			"Conclusions: To be specific, it produces the following contributions to the SPS and SEE communities: … It presents a most recent update and exhaustive collection of the research on this growing topic; … larger number of relevant studies were identified than ever …; The review scope first explicitly combined the simulation games and non-game simulators …; The existing SEE simulators were described with their individual capabilities …; An initial version of the recommendation checklist was proposed …"
S16	3 years	Yes	"Abstract: Our analysis reveals a field that is rapidly growing, both in volume and empirical focus, while becoming more mature."
			"Abstract: We identify signs of field maturity from the increase in: (i) the number of journal articles, (ii) the empirical models within the last two years, and (iii) the number of ecosystems studied."
			"Introduction: Our analysis identifies two main issues that inhibit the evolution of the field, …"
			"Introduction: … our analysis indicates that research has evolved outside the borders of the existing definitions …"
			"Analysis: We notice that the number of publications has been increasing since 2007."
			"Analysis: We notice that the tendency of increased publications along the years is also followed by a tendency of increased journal papers…."
			"Analysis: We note that there is an increase in the percentage of the groups 'report' and 'empirical model', while there is a decrease in 'tools or notation', 'qualitative or descriptive models', 'analytic models', and 'specific solution'."
			"Analysis: We find that there are some new concepts that appear in the literature, …"
			"Analysis: .. we identify some emerging concepts in the new literature …"
			"Discussion: we note that the field has been increasing significantly in volume …"
			"Summary: We document the evolution in the field by analyzing the literature as a time sequence …"
			"Summary: our analysis shapes the view of the field having evolved outside the existing definitions of software ecosystems and, thus, propose the update of the definition of software ecosystems."
S17	2 years	Yes	"Abstract: The findings show that quantitative studies and the theoretical underpinnings of agile ISD are lacking."
			"Abstract: In consequence, multiple research gaps that need further research attention are identified."
			"Discussion: This review extends findings of previous reviews (Hoisl and Sobernig, 2016, Gao et al., 2014, Leyh and C., 2012) by introducing new perspectives and by including contributions until early 2013."
			"Discussion: Furthermore, this review exposed promising research areas by presenting a systematic map on the focus of the studies and the employed research methods."
			"Discussion: To sum up, this review serves as foundation for future studies by opening up several research possibilities …"
S18	8 years	Yes	"Abstract: Key results include that in contrast to 1999–2009,"
			"Research Motivation: … although there have been reviews covering different aspects of agile GSD in the past, we have identified the research gap that there is no recent study focusing …"
			"Research Motivation: … This study is furthermore unique in the way that it effectively covers the usage of agile practices in GSD from agile origins (1999) to recent time (2016)."
			"Research Motivation: … we see particular value in our study by continuing the work of Jalali and Wohlin (S62; S63) to perform a complete and updated analysis in contrast to covering the whole time frame on our own"
			"Results: Compared to years 1999–2009 (S63), there is an increase in total paper count per year …"
			"Results: we can see a maturing research field with a shift from experience reports, …"
			"Results: a decrease in the usage of purely qualitative methods in favor of more mixed and quantitative approaches."
			"Results: there is a rise in literature reviews …"
			"Results: although overall the most frequently seen contribution (51%), have decreased drastically from 70% in 1999–2009 to 20% in 2010–2016 in favor of more rigorous contributions such as case study analysis …"
			"Results: For 2010–2016, although most studies still lack some details of the empirical context, …"
			"Results: In our newer studied period of 2010–2016, the customer has been even less frequently reported and remained unclear in 65% of relationships …"
			"Results: From 1999 to 2016, the standup meeting has been most frequently applied (70.5 cases), …"
			"Discussion: We have found a great variety of distribution scenarios in years 2010–2016 and were able to provide deeper insight compared to (S63), …"
			"Conclusion: The research contribution of this study lies in reporting the state of the art of agile practices in GSD for 1999–2016, where we almost exclusively found success stories."
			"Conclusion: We also saw several modifications and extensions of agile practice in use to deal with the challenges of GSD."
S19	3 years	No	"Abstract: This paper presents recommendations …. Recommendations are based on a previous systematic review, which was updated and expanded with 13 new empirical studies and more than 60 new empirical results."
			"Abstract: … it was found that interviews were one … The new evidence supports the same conclusion."
			""Main findings: The sensitivity analysis had been repeated in our updating SR … The final result is the same. …"
			"Main findings: … it was not possible to corroborate the similarity of (de Paula and Carneiro, 2016) and (Kitchenham and Charters, 2007) due to the fact that …"
			""Discussion: The newly obtained evidence points out in the same direction than the previously existing one."
			""Discussion: … the present study gathers more evidence for the effectiveness of interviews and laddering."
			"Discussion: The inferiority of sorting techniques is also corroborated by the present study …"
			"Discussion: The present study also confirms that the efficacy of interviews, laddering, …"
			"Discussion: … we did not find new evidence about a new technique (scaling) not considered before."

Table C.1 (continued)

ID	Publication timespan	Changes in conclusions?	Evidence
S20	6 years	Yes	"Discussion: … it is noticeable that only 3 of the 9 papers reviewed …. have contributed to the set of aggregations …" "Discussion: Quite the contrary, the same fact was already observed in the previous review …" "Conclusions: We already established in (Garritty et al., 2010, Leyh and C., 2012) that interviews were able to extract more information …" "Conclusions: … In the present work, we could not find any contradictory evidence … the results corroborate the superiority of interviews …" "Abstract: Compared the original study, the larger group of empirical studies of the UML through our analysis concentrates on comprehension" "Introduction: we mainly focus on the change and the development trend in the research of UML in recent years by means of extension into an SLR …" "Discussion: had a tendency to continue for future studies (this validate the prediction of the original research). …" "Discussion: The number of empirical studies focuses on UML metrics has presented a significant decreasing …" "Discussion: Not surprisingly, the type of diagram that is most frequently studied show a clear sequence: class diagrams …" "Discussion: This result confirms one of the points of the several related work and original study …" "Discussion: Although the original study (de Paula and Carneiro, 2016) has failed to show a direct statement, we could still derive the general trend …" "Conclusion: we had presented an extension of a systematic literature review for assessing the further development trend of the notations of the UML and the effectiveness in practical application …" "Conclusion: We also should note more studies are also necessary conducted with professional software engineers or practitioners to raise the generalizability of the results."

References

Alabool, H.M., Mahmood, A.K., 2013. Review on cloud service evaluation and selection methods. In: 2013 International Conference on Research and Innovation in Information Systems (ICRIIS' 13), pp. 61–66.

Alabool, H., Kamil, A., Arshad, N., Alarabiat, D., 2018. Cloud service evaluation method-based Multi-Criteria Decision-Making: A systematic literature review. Journal of Systems and Software 139 (May 2018), 161–188.

Ameller, D., Farré, C., Franch, X., Rufian, G., 2016. A Survey on Software Release Planning Models. In: 17th International Conference on Product-Focused Software Process Improvement (PROFES' 16), Lecture Notes in Computer Science, 10027, pp. 48–65.

Arshad, S., Usman, M., Ikram, N., 2012. Knowledge Management Practices in GSD: A Systematic Literature Review. In: 17th International Conference on Software Engineering Advances (ICSEA' 12), pp. 516–523.

Bayliss, H.R., Haddaway, N.R., Eales, J., Frampton, G.K., Jame, K.L., 2016. Updating and amending systematic reviews and systematic maps in environmental management. Environmental Evidence 5–20.

Beecham, S., Baddoo, N., Hall, T., Robinson, H., Sharp, H., 2008. Motivation in Software Engineering: A systematic literature review. Information and Software Technology 50 (9–10) 86–78.

Beninca, R., Huzita, E., Galdamez, E., Leal, G., Balancieri, R., Massago, Y., 2015. Knowledge management practices in GSD – a systematic literature review update. In: 17th International Conference on Enterprise Information Systems (ICEIS' 15), pp. 365–373.

Bezerra, R., da Silva, F., Santana, A., de Magalhães, C., Santos, R., 2015. Replication of Empirical Studies in Software Engineering: An Update of a SystematicMapping Study. In: Intl. Symp. on Emp. Softw. Eng. and Measurement., pp. 1–4.

Bornmann, L., 2014. How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. Res. Eval. 23, 166–173.

Boyle, E.A., Hainey, T., Connolly, T.M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., Pereira, J., 2016. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers & Education 94 (1), 178–192.

Budgen, D., Burn, A.J., Brereton, O.P., Kitchenham, B.A., Pretorius, R., 2011. Empirical evidence about the UML: a systematic literature review. Software Practice and Experience 41, 363–392.

Chung, M., Newberry, S.J., Ansari, M.T., Yu, W.W., Wu, H., Lee, J., Suttorp, M., Gaylor, J.M., Motala, A., Moher, D., Balk, E.M., Shekelle, P.G., 2012 Jun. Two methods provide similar signals for the need to update systematic reviews. J Clin Epidemiol. 65 (6), 660–668. doi:10.1016/j.jclinepi.2011.12.004. Epub 2012 Mar 29.

Connolly, T.C., Boyle, E.A., Hainey, T., McArthur, E., Boyle, J.M., 2012. A systematic literature review of empirical evidence on computer games and serious games. Computers & Education 59, 661e686.

Dantas, E., Perkusich, M., Dilorenzo, E., Santos, D., Hyggo, A., Perkusich, A., 2018. Effort Estimation in Agile Software Development: An Updated Review. International Journal of Software Engineering and Knowledge Engineering 28 (11n12), 1811–1831.

Dieste, O., Juristo, N. "Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques", IEEE Transactions on Software Engineering, Volume 37, Issue 2, p. 283–304, 2010.

de Paula, A.C.M., Carneiro, G.F., 2016. "Cloud Computing Adoption, Cost-benefit Relationship and Strategies for Selecting Providers: A Systematic Review". In: 11th International Conference on Evaluation of Novel Software Approaches to Software Engineering (ENASE' 16), pp. 27–39 2016.

de Paula, A.C.M., de Carneiro, G.F., 2016. A Systematic Literature Review on Cloud Computing Adoption and Migration. Evaluation of Novel Approaches to Software Engineering (ENASE' 16), 703.

Dieste, O., López, M., Ramos, F., 2008a. Updating a systematic review about selection of software requirements elicitation techniques. In: 11th Workshop on Requirements Engineering (WER' 08), pp. 96–103.

Dieste, O., Marta, T., Ramos, L.F., 2008b. Obtaining Well-Founded Practices about Elicitation Techniques by Means of an Update of a Previous Systematic Review. In: 20th International Conference on Software Engineering and Knowledge Engineering (SEKE' 08), pp. 769–772.

Dingsøyr, T., Nerur, S., Balijepally, V., Moe, N.B., 2012. A decade of agile methodologies: Towards explaining agile software development. Journal of Systems and Software 85 (6), 1213–1221.

Ferrari, F.C., Pizzoleto, A.V., Offutt, A.J., 2018. A systematic review of cost reduction techniques for mutation testing: preliminary results. In: 13th International Workshop on Mutation Analysis (Mutation' 18), pp. 1–10.

França, A.C.C., Gouveia, T.B., Santos, P.C.F., 2011. Motivation in software engineering: a systematic review update. In: 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE' 11), pp. 154–163.

Gao, C., Jiang, S., Rong, G., 2014. Software process simulation modeling: preliminary results from an updated systematic review. In: 2014 International Conference on Software and System Process (ICSSP' 14), pp. 50–54.

Garner, P., Hopewell, S., Chandler J, J., MacLehose, H., et al., 2016. When and how to update systematic reviews: consensus and checklist. BMJ 354, i3507. doi:10.1136/bmj.i3507, pp. 1-10, 2016.

Garousi, V., Fernandes, J.M., 2016. Highly-cited papers in software engineering: The top-100. Information & Software Technology 71, 108–128.

Garritty, C., Tsertsvadze, A., Tricco, A.C., Sampson, M., Moher, D., 2010. Updating Systematic Reviews: An International Survey. PLoS ONE 5 (4), e9914.

Ghazi, A.N., Petersen, K., Reddy, S.S.V.R., Nekkanti, H., 2018. Survey research in software engineering: problems and mitigation strategies, 6. IEEE Access, p. 1.

Guo, M., Zhang, C., Wang, F., 2017. What is the Further Evidence about UML? - A Systematic Literature Review. 24th Asia-Pacific Software Engineering Conference Workshops (APSECW' 17), pp. 106–113.

Higgins, J.P.T., Green, S., Scholten, R.J.P.M. "Chapter 3: Maintaining reviews: updates, amendments and feedback", Higgins JPT, Green S, Version 5.1.0 [updated March 2011]. Cochrane Collaboration, 2011.

Hoisl, B., Sobernig, S., 2016. Open-Source Development Tools for Domain-Specific Modeling: Results from a Systematic Literature Review. In: 49th Hawaii International Conference on System Sciences (HICSS' 16), pp. 5001–5010.

Hopewell, S., Loudon, K., Clarke, M., Moher, D., Scholten, R., Eisinga, A., French, S., 2008. A decision tool for updating Cochrane reviews. Cochrane Centre UK. Cochrane Opportunities Fund. https://dspace.stir.ac.uk/handle/1893/24190#.XU6S_vlzapo.

Hummel, M., 2014. State-of-the-Art: A Systematic Literature Review on Agile Information Systems Development. In: 47th Hawaii International Conference on System Sciences (HICSS' 14), pp. 4712–4721.

Jalali, S., Wohlin, C., 2012. Global software engineering and agile practices: a systematic review. Journal of Software: Evolution and Process 24 (6), 643–659.

Jalali, S., Wohlin, C, C, 2010. Agile practices in global software engineering – A systematic map. In: 5th IEEE International Conference on Global Software Engineering (ICGSE' 10). IEEE, pp. 45–54.

Jiang, S., Zhang, H., Gao, C., Shao, D., Rong, G., 2015. Process simulation for software engineering education. In: International Conference on Software and System Process (ICSSP' 15), pp. 147–156.

Kitchenham, B.A., Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, School of Computer Science and Mathematics. Keele University.

Kitchenham, B.A., Dybå, T., Jørgensen, M., 2004. Evidence-Based Software Engineering. In: 26th International Conference on Software Engineering (ICSE' 04), pp. 273–281 B.

Kitchenham, B.A., Mendes, E., Travassos, G.H., 2006. A Systematic Review of Cross- vs. Within-company Cost Estimation Studies. In: 11th Annual Conference on Evaluation & Assessment in Software Engineering (EASE' 06), pp. 1–10.

Kitchenham, B.A., Mendes, E., Travassos, G.H., 2007. Cross versus within-company cost estimation studies: A systematic review. IEEE Transactions on Software Engineering 33 (5), 316–329.

Kitchenham, B.A., Brereton, O.P., Budgen, D., Turner, M., Linkman, S., 2009. Systematic literature reviews in software engineering – a systematic literature review. Information and Software Technology 51 (1), 7–15.

Kitchenham, B.A., Pretorius, R., Budgen, D., Brereton, O.P., Turner, M., Niazi, M., Linkman, Stephen, 2010. Literature reviews in software engineering – a tertiary study. Information and Software Technology 52 (8), 792–805.

Kitchenham, B.A., 2004. Procedures for Performing Systematic Reviews. Technical Report TR/SE-0401 ISSN:1353–7776, School of Computer Science and Mathematics. Keele University.

Leyh, C., C., 2012. Critical Success Factors for ERP System Implementation Projects: A Literature Review. In: Møller, C., Chaudhry, S. (Eds.), Advances in Enterprise Information Systems II. CRC Press/Balkema, Leiden, The Netherlands, pp. 45–56.

Leyh, C., Sander, P., 2015. Critical success factors for ERP system implementation projects: An update of literature reviews. In: Enterprise Systems. Strategic, Organizational, and Technological Dimensions. Lecture Notes in Business Information Processing, 198. Springer, pp. 45–67.

Linåker, J., Sulaman, S.M., Maiani de Mello, R., Höst, M., 2015. Guidelines for Conducting Surveys in Software Engineering. Technical Report. Lund University, Sweden.

Manikas, K., Hansen, K.M., 2002. Software ecosystems – a systematic literature review. Journal of Systems and Software 86 (5), 1294–1306.

Manikas, K., 2016. Revisiting software ecosystems Research: A longitudinal literature study. Journal of Systems and Software 117 (C), 84–103.

Maplesden, D., Tempero, E., Hosking, J., Grundy, J.G., 2015. Performance analysis for object-oriented software: A systematic mapping. IEEE Transaction on Software Engineering 41 (7).

Matos, O., Conte, T., Mendes, E., 2014. Is there a place for qualitative studies when identifying effort predictors?: a case in web effort estimation. In: 18th International Conference on Evaluation and Assessment in Software Engineering (EASE' 14), Article 40, pp. 1–10.

Mendes, E., Kalinowski, M., Martins, D., Ferrucci, F., Sarro, F., 2014. Cross vs. within–company cost estimation studies revisited: An extended systematic review. 18th Annual Conference on Evaluation & Assessment in Software Engineering (EASE' 14). ACM 12:1–12:10.

Mendes, E., Felizardo, K., Wohlin, C., Kalinowski, M., 2019. Search Strategy to Update Systematic Literature Reviews in Software Engineering. In: Proceedings of EUROMICRO 2019.

Moher, D., Tsertsvadze, A., Tricco, A.C., Eccles, M., Grimshaw, J., et al., 2007. A systematic review identified few methods and strategies describing when and how to update systematic reviews. Journal of Clinical Epidemiology 60 (1095) e1091–1095.e1011.

Moher, D., Tsertsvadze, A., Tricco, A.C., Eccles, M., Grimshaw J., et al., "When and how to update systematic reviews. Cochrane Database Systematic Review: MR000023, 2008.

Nair, S., de la Vara, J.L., Sabetzadeh, M., Briand, L., 2013. Classification, structuring and assessment of evidence for safety – a systematic literature review. In: 6th IEEE International Conference on Software Testing, Verification and Validation (ICST' 13), pp. 94–103.

Nair, S., de la Vara, J.L., Sabetzadeh, M., Briand, L., 2014. An extended systematic literature review on provision of evidence for safety certification. Information and Software Technology (IST) 56 (7), 689–717.

Nekkanti, H., Reddy, S.S.V.R. "Surveys in Software Engineering A Systematic Literature Review and Interview Study", Masters thesis number MSSE-2016-26, 2016, Masters of Science in Software Engineering, http://www.diva-portal.org/smash/get/diva2:1045814/FULLTEXT02.

Nepomuceno, V., Soares, S., 2018. Maintaining Systematic Literature Reviews: Benefits and Drawbacks. In: ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM' 18), pp. 1–4.

Oliveira, J.A., Vargas, M., Rodrigues, R., 2018. SOA Reuse: Systematic Literature Review Updating and Research Directions. In: 14th Brazilian Symposium on Information Systems (SBSI' 18), Article 71, pp. 1–8.

Pattanittum, P., Laopaiboon, M., Moher, D., Lumbiganon, P., Ngamjarus, C., 2012. A Comparison of Statistical Methods for Identifying Out-of-Date Systematic Reviews. PLoS ONE 7 (11), e48894.

Pizzoleto, A.V., Ferrari, F.C., Offutt, J., Fernandes, L., Ribeiro, M., 2019. A systematic literature review of techniques and metrics to reduce the cost of mutation testing. Journal of Systems and Software 157 (November 2019), 110388.

Riaz, M., Mendes, E., Tempero, E., 2009. A systematic review of software maintainability prediction and metrics. In: 3rd International Symposium on Empirical Software Engineering and Measurement (ESEM' 09), pp. 367–377.

Riaz, M., 2012. Maintainability prediction of relational database-driven applications: A systematic review. In: 16th International Conference on Evaluation & Assessment in Software Engineering (EASE' 12), pp. 263–272.

Sampson, M., Shojania, K.G., McGowan, J., Daniel, R., Rader, T., Iansavichene, A.E., Ji, J., Ansari, M.T., Moher, D., 2008, Aug. Surveillance search techniques identified the need to update systematic reviews. J Clin Epidemiol. 61 (8), 755–762. doi:10.1016/j.jclinepi.2007.10.003, Epub 2008 Feb 14.

Sobernig, S., Hoisl, B., Strembeck, M., 2016. Extracting reusable design decisions for UML-based domain-specific languages: A multi-method study. Journal of Systems and Software 113 (2016), 140–172.

Soomro, A.B., Salleh, N., Nordin, A., 2015. How personality traits are interrelated with team climate and team performance in software engineering? A preliminary study. In: 9th Malaysian Software Engineering Conference (MySEC' 15), pp. 259–265.

Sulayman, M., Mendes, E., 2011. An extended systematic review of software process improvement in small and medium web companies. In: 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE' 11), pp. 134–143.

Svahnberg, M., Gorschek, T., Feldt, R., Torkar, R., Bin Saleem, S., Shafique, M.U., 2010. A systematic review on strategic release planning models. Information and Software Technology 52 (3), 237–248 2010.

Takwoingi, Y., Hopewell, S., Tovey, D., Sutton, A.J., 2013 Dec 13. A multicomponent decision tool for prioritising the updating of systematic reviews. BMJ 347, f7191. doi:10.1136/bmj.f7191.

Usman, M., Mendes, E., Weidt, F., Britto, R., 2014. Effort estimation in agile software development: A systematic literature review. In: 10th International Conference on Predictive Models in Software Engineering – (PROMISE' 14), pp. 82–91.

Vallona, R., Estácio, B.J.S., Prikladnickic, R., Grechenig, T., 2018. Systematic literature review on agile practices in global software development". Information and Software Technology 96 (April 2018), 161–180.

Wagner, S., Mendez, D., Felderer, M., Kalinowski, M., 2020. Challenges in Survey Research. In: In Contemporary Empirical Methods in Software Engineering. Springer, pp. 95–128.

Zhang, H., Kitchenham, B.A., Pfahl, D., 2008. Reflections on 10 years of software process simulation modeling: A systematic review. In: 9th International Conference on Software Process (ICSP 2008), pp. 345–356.

Zhang, H., Kitchenham, B.A., Pfahl, D., 2010. Software Process Simulation Modeling: An Extended Systematic Review. In: New Modeling Concepts for Today's Software Processes (ICSP' 10), 6195. Lecture Notes in Computer Science, pp. 309–320.

Zhang, H., Babar, M.A., Tell, P., 2011. Identifying relevant studies in software engineering. Information and Software Technology 53 (6), 625–637.

Zhou, X., Jin, Y., Zhang, H., Li, S., Huang, X., 2016. A Map of Threats to Validity of Systematic Literature Reviews in Software Engineering. In: Proceedings of 23rd Asia-Pacific Software Engineering Conference (APSEC).

Emilia Mendes PhD 1999 Computer Science; MSc. 1992 Computer Science; BSc 1986 Computer Science; Full Professor in Computer Science at the Blekinge Institute of Technology (Sweden), Tekes-funded Finnish Distinguished Professor at the University of Oulu (Finland) (2015–2018). Main research areas of interest are empirical software engineering, Value-based Software Engineering, and the use of machine learning techniques to contexts such as healthcare, and sustainability. She has published widely and over 200 refereed papers, plus two books as solo author – both in the area of cost estimation. She has been on the editorial board of TSE, SQJ and EMSE, and is currently on IST's editorial board. Further details are available here: https://scholar.google.se/citations?user=Mz4IbXgAAAAJ&hl=en&oi=ao.

Claes Wohlin is a professor of software engineering at Blekinge Institute of Technology, Sweden. He served as Dean for the Faculty of Computing for five years (2014–2018) and Pro-Vice-Chancellor for six years (2004–2010). Claes Wohlin has previously held professor chairs at the universities in Lund and Linköping. He received a PhD in communication systems from Lund University. In 2019, he was appointed Emeritus Editor for the Journal of Information and Software Technology after having served as Editor-in-Chief for 14 years and Editor for three years. Claes Wohlin received Telenor's Nordic Research Prize in 2004 for his achievements in software engineering and the improvement of reliability for telecommunication systems. His-research interests include empirical methods in software engineering, global software engineering and software quality. In 2011, he was elected member of the Royal Swedish Academy of Engineering Sciences.

Katia Romero Felizardo is Associate Professor at the Federal Technological University of Paraná (UTFPR - Cornélio Procópio). She coordinates the Graduate Program in Computing and supervises Masters and PhD students in topics mainly related to Evidence Based Software Engineering (EBSE). Her main research interests focus upon secondary studies: systematic literature reviews and systematic mappings. She currently leads a research project funded by CNPq. She has published 30+ papers in peer-reviewed national and international conferences and journals, mainly in the area of EBSE. A list of her main publications can be found here: https://scholar.google.com/citations?hl=en&user=YByP4IQAAAAJ.

Marcos Kalinowski is a professor of software engineering at the Informatics Department of the Pontifical Catholic University of Rio de Janeiro, PUC-Rio, Brazil. Serves as the coordinator of the graduate program in Informatics (2020–2022). He received a PhD in computer and systems engineering from the Federal University of Rio de Janeiro. Serves the co-chair of the 'In Practice' track of the Journal of Systems and Software (since 2020), were he is also active at the Editorial Board (since 2018). His-research interests include empirical methods in software engineering, software quality, and software engineering for artificial intelligence and digital transformation. He has been active in international research collaborations. In 2015, he was voted member of the International Software Engineering Research Network (ISERN).