Information and Software Technology 112 (2019) 48–50

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

A critical appraisal tool for systematic literature reviews in software engineering☆

Nauman bin Ali*, Muhammad Usman

Blekinge Institute of Technology, Karlskrona, Sweden

ARTICLE INFO

Keywords:
Systematic literature reviews
Quality assessment
Software engineering
Critical appraisal tools
AMSTAR

ABSTRACT

Context: Methodological research on systematic literature reviews (SLRs) in Software Engineering (SE) has so far focused on developing and evaluating guidelines for conducting systematic reviews. However, the support for quality assessment of completed SLRs has not received the same level of attention.
Objective: To raise awareness of the need for a critical appraisal tool (CAT) for assessing the quality of SLRs in SE. To initiate a community-based effort towards the development of such a tool.
Method: We reviewed the literature on the quality assessment of SLRs to identify the frequently used CATs in SE and other fields. *Results:* We identified that the CATs currently used is SE were borrowed from medicine, but have not kept pace with substantial advancements in the field of medicine.
Conclusion: In this paper, we have argued the need for a CAT for quality appraisal of SLRs in SE. We have also identified a tool that has the potential for application in SE. Furthermore, we have presented our approach for adapting this state-of-the-art CAT for assessing SLRs in SE.

1. Introduction

Inspired by medicine, evidence-based software engineering (EBSE) promotes the use of systematic literature reviews (SLRs) to systematically identify, evaluate and synthesize research on a topic of interest [1]. Since the introduction of SLRs in Software Engineering (SE), the rate of papers reporting SLRs in SE[1] has been continually increasing (see Fig. 1).

However, several recent in-depth evaluations of published SLRs have identified serious flaws regarding their quality. For example, issues related to: (a) the reporting quality of procedures and outcomes [2], (b) the reliability of search [3], and (c) lack of synthesis or the use of inappropriate synthesis methods [4] in SLRs. Such issues raise questions about the credibility of SLRs.

Most researchers in SE have used the four questions adopted by Kitchenham et al. [5] (items *a* to *d* in Table 1) for quality assessment of SLRs. These questions are insufficient to reveal important limitations in an SLR as demonstrated by the above-listed studies [2–4].

Fig. 2 helps to understand the role of guidelines and distinguish the purpose of critical appraisal tools (CAT). The guidelines for planning and conducting an SLR enable a research team to plan and execute a review that follows a rigorous process [1,5]. Similarly, the reporting guidelines help the researchers to communicate the design and execution of an SLR to the readers [1]. More recently, there are new reporting guidelines that are intended to improve the usefulness of the results of an SLR for education and practice [6].

On the other hand, the role of critical appraisal tools is to facilitate a reader to analytically assess the credibility of a completed SLR. Such an assessment considers both the reporting quality, e.g., *"Are the review's inclusion and exclusion criteria described?"*, and the risk of bias assessment in the design and execution of the SLR, e.g., *"Are the review's inclusion and exclusion criteria appropriate?"*.

As the number of SLRs is increasing, the need for tools to assess the quality of an SLR without having to replicate the study is becoming more evident. Such a CAT will help to sustain and improve the credibility of SLRs as an effective means for decision-support in SE. It will enable the readers of SLRs to differentiate between good quality SLRs from the ones that did not follow a rigorous and comprehensive approach.

In this paper, we seek to raise awareness of the need for a critical appraisal instrument and have introduced a candidate solution for this task. The work presented in this paper has the potential to have a

* Corresponding author.
E-mail addresses: nauman.ali@bth.se (N.b. Ali), muu@bth.se (M. Usman).

[1] (Search string used in Scopus to identify SLRs published in computing TITLE-ABS-KEY("systematic review" OR "systematic literature review") AND PUBYEAR < 2019 AND (LIMIT-TO (SUBJAREA,"COMP ")).

https://doi.org/10.1016/j.infsof.2019.04.006
Received 5 November 2018; Received in revised form 5 April 2019; Accepted 12 April 2019
Available online 15 April 2019

Table 1
AMSTAR-2, and DARE quality criteria used to appraise SLRs.

DARE - *Note: Fulfilling items a, b and e, and either c or d is mandatory for an SLR to be included in the DARE database of SLRs.*	
a. Were inclusion/exclusion criteria reported?	d. Are sufficient details about the individual included studies presented?
b. Was the search adequate?	e. Were the included studies synthesised?
c. Was the quality of the included studies assessed?	
AMSTAR -2 - *Note: Items marked with an asterisk (*) are not applicable for the appraisal of SMSs.*	
1. "Did the research questions and inclusion criteria for the review include the components of PICO?"	
2. "Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?"	
3. "Did the review authors explain their selection of the study designs for inclusion in the review?"	
4. "Did the review authors use a comprehensive literature search strategy?"	
5. "Did the review authors perform study selection in duplicate?"	
6. "Did the review authors perform data extraction in duplicate?"	
7. "Did the review authors provide a list of excluded studies and justify the exclusions?"	
8. "Did the review authors describe the included studies in adequate detail?"	
9.* "Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?"	
10. "Did the review authors report on the sources of funding for the studies included in the review?"	
11.* "If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?"	
12.* "If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?"	
13.* "Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?"	
14.* "Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?"	
15.* "If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?"	
16. "Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?"	

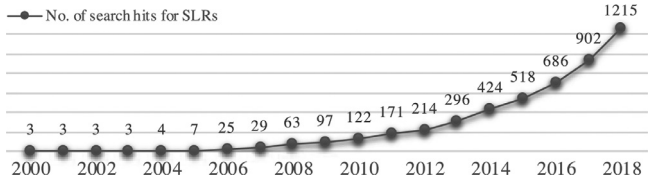

Fig. 1. The increasing number of SLRs in computing since 2004.

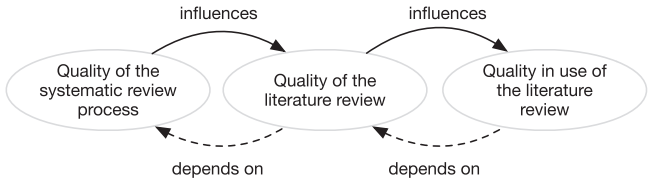

Fig. 2. A view of quality on the various stages of an SLR.

profound impact on SE research, since it is useful for two very common scenarios: (1) to assess the quality of an SLR as a referee/reader, and (2) to synthesize the results of several SLRs on the same topic and to understand the reasons for any differences between their results.

The remainder of the paper is structured as follows: Section 2 explains the need for a CAT for SLRs in SE. Section 3 presents a state-of-the-art CAT. In Sections 4 and 5, we briefly propose an approach to customize and validate the tool for SE. Section 6 concludes the paper.

2. Need for a CAT for the quality assessment of SLRs in SE

Since 2004, when the first guidelines for SLRs in SE were introduced, several improvements have been made to the guidelines for conducting and reporting SLRs in SE [1]. However, the appraisal tools for SLRs in SE have not received much attention. Researchers in the SE field have continued to rely on a subset of questions identified by Kitchenham et al. [5] from the field of evidence-based medicine in the year 2004. The commonly-used interpretation of the DARE[2] criteria in SE does not even consider if there is a synthesis performed in a review. This explains to some extent why some of the limitations in the quality of SLRs e.g. poor reporting quality [2], lack of an adequate search strategy [3] and the lack of synthesis [4] cannot be sufficiently revealed with the CATs currently used in SE.

In the meantime, realizing the importance of CATs to assess the quality of completed systematic reviews, researchers in other disciplines have further developed these tools. A review of evidence-based medicine literature reveals that one tool that stands out for the degree of validation and application is AMSTAR (A MeaSurement Tool to Assess systematic Reviews) [7]. AMSTAR was developed based on a scoping review of the then available rating instruments. The review identified several over-lapping appraisal items, which were combined into 11 AMSTAR appraisal items using factor analysis [7]. After pilot testing, the original AMSTAR was validated externally as well [8]. AMSTAR[3] has since then been used and validated extensively [8,9].

3. Candidate CAT for quality assessment of SLRs in SE

Recently, the designers of AMSTAR have proposed a revision of the tool (AMSTAR-2 [8]). The revision is based on community feedback collected through different channels such as published reports of its application, the AMSTAR website,[4] surveys of AMSTAR users, and the experience of participants in AMSTAR workshops. The team that has revised the tool includes designers of the original instrument and two designers of another instrument ROBIS (Risk Of Bias In Systematic reviews). ROBIS[5] is a relatively new instrument and is designed to support reviewers in assessing the risk of bias in completed SLRs.

AMSTAR-2 can be used to appraise SLRs that may include both randomized or non-randomized studies. AMSTAR-2 has a more detailed assessment of the risk of bias in SLRs due to the primary studies included, and how the review authors have dealt with such bias when interpreting review results. AMSTAR-2 consists of 16 items (see Table 1), and each item has detailed response options to guide users to make the appropriate judgement (see complete AMSTAR-2[4] for details). The initial

[2] The CRD's Database of Abstracts of Reviews of Effects (DARE) https://www.crd.york.ac.uk/CRDWeb/AboutPage.asp.

[3] The AMSTAR paper [7] had 2958 citations on February 13, 2018.

[4] AMSTAR https://amstar.ca/.

[5] ROBIS https://www.bristol.ac.uk/population-health-sciences/projects/robis/robis-tool/.

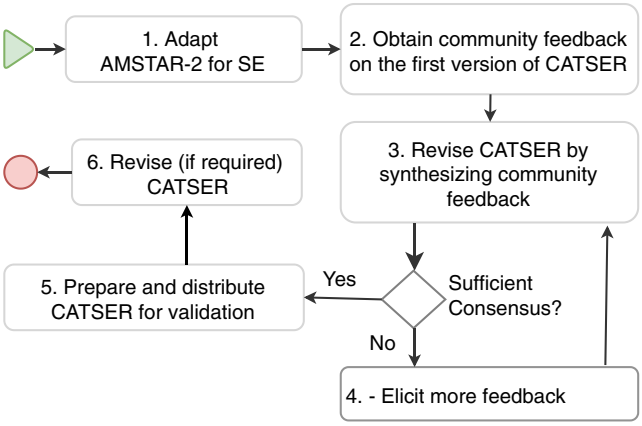

Fig. 3. Approach for adapting AMSTAR-2 for SE.

evaluation of AMSTAR-2, by having multiple raters use the tool, has shown moderate to good agreement for most items in the tool [8].

AMSTAR-2 has several advantages over the DARE criteria commonly used in SE. DARE is not a CAT per se; it is intended to provide the criteria that SLRs should meet to be included in the CRD's database of SLRs. In SE, only four of the five items of DARE (items *a* to *d* in Table 1) have often been used [5]. Apart from item *b* the formulation of DARE items only captures the reporting quality in SLRs (cf. [5]), e.g., see item *a* about reporting of the selection criteria. Furthermore, many of the items in AMSTAR-2 (e.g. items 1, 5, 6, 7, 10, 14, 15, and 16) which capture the quality of an SLR are not covered by the DARE criteria.

In this study, we have identified AMSTAR-2 as a candidate CAT that can be adapted for SE. The approach we will use in developing and validating CATSER (a Critical Appraisal Tool for SE systematic Reviews based on AMSTAR-2) is described in the following sections and depicted in Fig. 3.

4. A proposed approach for adapting AMSTAR-2 for SE

We propose to first adapt AMSTAR-2 for SE by reviewing its items and response options for their relevance to SE using the recommendations in the EBSE literature (e.g., [1,10]). In the next phase, we will involve the SE research community for the further evolution of CATSER. We will organize workshops at the prominent SE venues (e.g., the international symposium on empirical software engineering and measurement (ESEM)[6]). Furthermore, a web-based forum will be set up to collect feedback from the wider community.

We have reviewed the relevance of AMSTAR-2 items for SE systematic secondary studies (systematic mapping studies (SMS) [1] and SLRs). Out of the 16 items in AMSTAR-2, we consider 10 items (see Table 1) relevant for the critical appraisal of both SLRs and SMSs. These 10 items cover the fundamental aspects (e.g., protocol development, systematic search, study selection, and data extraction processes) necessary for the reliability of both SLRs and SMSs.

SMSs do not include a thorough synthesis and detailed quality assessment of the included primary studies [1]. Therefore, we consider the remaining six items regarding synthesis and meta-analysis as only relevant for SLRs.

The response options in AMSTAR-2 are formulated for the medical discipline, and these will require adaptation for SE. For this purpose, we will use the latest guidelines for designing, reporting, conducting and validating systematic secondary studies in SE [1–3,10].

[6] http://www.esem-conferences.org/.

5. A proposed approach for validating CATSER

We plan to validate CATSER by using it to appraise a set of SLRs using reviewers beyond those who will be involved in the adaptation of AMSTAR-2 for SE. We will allocate a small sample of randomly selected SLRs to the reviewers. Using the results of individually appraised SLRs with CATSER, we plan to compute the inter-rater reliability of CATSER.

Another aspect of the evaluation of CATSER will focus on its usefulness to identify significant flaws in an SLR. In the future, we will compare the assessment of SLRs using CATSER and the commonly-used interpretation of DARE in SE.

The long-term validation of such instruments depends on how widely they are accepted and used by the community. We hope to initiate a community effort in SE to adapt, validate and mature CATSER (which will leverage the strengths of AMSTAR-2).

6. Conclusion

By comparing the state-of-the-art tools in medicine with the frequently used CATs in SE, and based on the recent evaluations of the quality of SLRs, we identified and emphasized the need for further research on CATs for SLRs in SE. We have also identified a candidate CAT and proposed an approach to adapt it for the needs of SE with the involvement of the SE research community.

This approach will not only improve the quality of the tool, but ensure community buy-in and thus increase the likelihood of adoption of the tool. Given the continued interest in SLRs in SE, we contend that this work has a potentially significant impact on research. It will help to improve and sustain the credibility of SLRs in SE.

Acknowledgment

The authors would like to thank Prof. Claes Wohlin for providing feedback on the paper. This work has been supported by a research grant for the VITS project (reference number 20180127) by the Knowledge Foundation in Sweden and by ELLIIT, a Strategic Area within IT and Mobile Communications, funded by the Swedish Government.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] B.A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, Chapman & Hall/CRC, 2015.
[2] D. Budgen, P. Brereton, S. Drummond, N. Williams, Reporting systematic reviews: some lessons from a tertiary study, Inf. Softw. Technol. 95 (2018) 62–74.
[3] N.B. Ali, M. Usman, Reliability of search in systematic reviews: towards a quality assessment framework for the automated-search strategy, Inf. Softw. Technol. 99 (2018) 133–147.
[4] D.S. Cruzes, T. Dybå, Research synthesis in software engineering: a tertiary study, Inf. Softw. Technol. 53 (5) (2011) 440–455.
[5] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, S. Linkman, Systematic literature reviews in software engineering - a tertiary study, Inf. Softw. Technol. 52 (8) (2010) 792–805.
[6] B. Cartaxo, G. Pinto, S. Soares, Towards a model to transfer knowledge from software engineering research to practice, Inf. Softw. Technol. 97 (2018) 80–82.
[7] B.J. Shea, J.M. Grimshaw, G.A. Wells, M. Boers, N. Andersson, C. Hamel, A.C. Porter, P. Tugwell, D. Moher, L.M. Bouter, Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews, BMC Med. Res. Methodol. 7 (1) (2007) 10.
[8] B.J. Shea, B.C. Reeves, G. Wells, M. Thuku, C. Hamel, J. Moran, D. Moher, P. Tugwell, V. Welch, E. Kristjansson, et al., AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both, BMJ 358 (2017) j4008.
[9] B.J. Shea, L.M. Bouter, J. Peterson, M. Boers, N. Andersson, Z. Ortiz, T. Ramsay, A. Bai, V.K. Shukla, J.M. Grimshaw, External validation of a measurement tool to assess systematic reviews (AMSTAR), PLoS One 2 (12) (2007) e1350.
[10] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, A. Chatzigeorgiou, Identifying, categorizing and mitigating threats to validity in software engineering secondary studies, Inf. Softw. Technol. 106 (2019) 201–230.