Evaluating strategies for study selection in systematic literature studies

Nauman Bin Ali
Blekinge Institute of Technology
37179 Karlskrona, Sweden
nauman.ali@bth.se

Kai Petersen
Blekinge Institute of Technology
37179 Karlskrona, Sweden
kai.petersen@bth.se

ABSTRACT

Context: The study selection process is critical to improve the reliability of secondary studies. *Goal*: To evaluate the selection strategies commonly employed in secondary studies in software engineering. *Method*: Building on these strategies, a study selection process was formulated and evaluated in a systematic review. *Results*: The selection process used a more inclusive strategy than the one typically used in secondary studies, which led to additional relevant articles. *Conclusions*: The results indicates that a good-enough sample could be obtained by following a less inclusive but more efficient strategy, if the articles identified as relevant for the study are a representative sample of the population, and there is a homogeneity of results and quality of the articles.

Categories and Subject Descriptors

D.2.0 [**Software Engineering**]: General

General Terms

Theory

Keywords

systematic review, inclusion and exclusion, study selection

1. INTRODUCTION

Secondary studies [3] [6] are used to explore a variety of topics in software engineering with an aim to answer a research question by conducting an *"exhaustive"* search for relevant literature [3]. Starting with a large set of potentially relevant studies, reviewers rely on several steps of selection, first by reading titles and abstracts, followed by full-text reading for quality assessment [3]. Thus, the reliability of a secondary study is highly dependent on the repeatability of the selection process [8].

Kitchenham and Brereton [4] have aggregated the research on the process of conducting systematic literature studies. The research focusing on the selection of studies has two complementary themes: (1) Using textual analysis tools to assist the selection process e.g. by visualization of citations and content maps to identify clusters of similar studies [2]. Existing work has shown feasibility of the approach and that it can reduce the effort spent on selection [7], but a thorough evaluation of the approach is still required [4]; (2) Making the inclusion/exclusion more systematic and identifying different strategies used in the selection process e.g. how disagreements in the selection of papers are handled [5].

Building upon the second theme, in this study, we contribute to the secondary study guidelines by providing an example combination of existing strategies, and an evaluation of their implication on the outcomes of the study. This study also provides a structured way to approach the resolution of disagreements and supports an informed decision-making considering the effort spent and the value achieved by different strategies.

The remainder of the paper is structured as follows: Section 2 describes the research method, Section 3 presents the study selection process, Section 4 presents the results and Section 5 concludes the paper.

ESEM'14, September 18–19, 2014, Torino, Italy.

http://dx.doi.org/10.1145/2652524.2652557

2. RESEARCH METHOD

The following three main sets of strategies were identified in literature [5]:

- *Formulate criteria objectively and evaluate objectivity by calculating the agreement level between reviewers*
- *Specify the decision rules to determine whether an article is included, excluded or there needs to be further investigation to decide about the selection result*
- *Specify rules to aid reviewers in resolving uncertainties and disagreements e.g. consulting additional information, seek an additional reviewers' input, voting, or discussion*

Individually these strategies are insufficient to ensure repeatability of the process however none of the existing reviews reported them together [5]. This raised two open questions i.e. which strategies should be combined and what order should they be combined in, to get higher effectiveness and efficiency in the selection process? To provide a partial answer to these more general questions we devised a study selection process (see Section 3, which builds on [5]) that was used to answer the following research question (RQ): *How effective and efficient are the employed selection strategies in including/excluding articles?*

By effectiveness, we refer to the ability to include relevant articles and to avoid their exclusion. Efficiency was measured in terms of the effort (time spent) when using the process. The RQ was answered by using the selection pro-

cess in a systematic review of software process simulation literature.

Threats to validity: The selection process has been used in only one systematic review. Thus, further replications are required to derive generalisable conclusions. The analysis in this study is based on the assumption that adaptive reading depth [6] produced as good results as full-text reading.

3. STUDY SELECTION PROCESS

An overview of the selection process used in this study is shown in Figure 1. To improve the quality of a secondary study, it is recommended that study selection is guided by already decided upon criteria in the study protocol [3]. As a first step, the criteria were specified as questions that can be objectively answered.

Once the criteria have been updated based on the review, the next step is to employ what we referred to as *"think-aloud protocol"*. In this step, reviewers apply the selection criteria on randomly selected articles from the pool of potentially relevant articles. A reviewer, while applying the criteria, also expresses their thoughts and rationales for their choice, which helps to clarify any ambiguities and unintended interpretations. This contributes to improving the internal consistency of studies.

The next step is to pilot the selection criteria [3] on a randomly selected subset of articles independently by the reviewers. Calculating inter-rater agreement values will indicate the objectivity of the criteria. If the results show disagreements, discussing these articles will help to clarify the ambiguities.

Once the reviewers are satisfied with the level of agreement, we apply the selection criteria on articles independently. Calculating the inter-rater agreement again should show confidence in the repeatability of the study. However, even with an acceptable level of inter-rater agreement we have to consciously decide how disagreements will be handled.

Use of the three possible labels for classifying an article by each reviewer i.e. relevant, irrelevant or uncertain results in a number of possible combinations of disagreements depending on the number of reviewers. The possible outcomes of the selection process with two reviewers with the above three possible classifications for each article are presented in Table 1. Unlike the list of possibilities identified from literature [5], this is a comprehensive list of categories.

Table 1: Different scenarios from study selection.

		Reviewer 2		
		Relevant	Uncertain	Irrelevant
Reviewer 1	Relevant	A	B	D
	Uncertain	B	C	E
	Irrelevant	D	E	F

This structured way of analysing the disagreements helps to see the level of disagreement and use this knowledge to focus additional effort for further analysis where it is needed. Moreover, the documentation of this systematic approach will improve the repeatability of studies as both the decision choices and the rationales will be available.

Existing decision rules used in secondary studies [5] make decisions like "if all reviewers agree that the paper should be excluded then it is excluded, otherwise it is included" would mean excluding articles in category *'F'*. However, such decision rules ignore the information that is now available about other articles e.g. articles in category *'E'* are most likely irrelevant and should not be taken for full-text reading.

An inclusive strategy will be to review and discuss all the articles in categories *'D'* and *'E'* to re-categorize such articles into categories *'A'*, *'C'* or *'F'*. Articles in category *'A'* and *'B'* should be included for full-text reading given the strong indication of relevance and articles in category F should be excluded.

Articles ending up in the uncertain category *'C'* could potentially be useful, but the title and abstract does not provide sufficient information to make an informed decision. To further investigate these articles in category *'C'*, adaptive reading depth can be used [6]. The outline of three step process used: read the introduction of the article to make a decision; if a decision is not reached read the conclusion of the article; if it is still unclear, search for the keywords and evaluate their usage to describe the context of the study in the article; if a decision is not reached mark the article as uncertain.

Table 1 of decision possibilities and their implications can be generalized for more than two reviewers. As an example, category *'A'* is reformulated to *if all reviewers classified an article as "relevant" implies that it should be included*; category *'B'* to *if more reviewers classified an article as "relevant" or "uncertain" than "irrelevant" implies an indication of relevance of the article*; etc.

Each reviewer applies the criteria individually, and logs not only the decision but also what stage of the adaptive reading process the decision was made. This information helps to discuss the disagreements and retrace the decision rationale faster. For inclusiveness, any articles where there is no consensus to exclude or include along with articles marked as uncertain are taken further for full-text reading.

4. RESULTS

A pilot study of our proposed selection process was conducted in the context of a systematic literature review aimed to aggregate evidence of the usefulness of software process simulation modelling. The main research question for the systematic literature review was: *What evidence has been reported that software process simulation models achieve their purposes in real-world settings?*

A search string was developed with relevant keywords and used in eight digital databases and one search engine. As the subject is well researched, we had a relatively large set of ***1906*** potentially relevant articles after the initial screening. The large number of search results and involvement of three reviewers (in the selection of studies) provided a good setting to apply the proposed selection process.

The initial protocol which was developed by the first two reviewers and reviewed by the third reviewer. It was decided that while applying the criteria only three possible outcomes will be used by each reviewer: relevant, irrelevant or uncertain (i.e. need more information than the title and abstract). The criteria were specified very objectively and the level of details of articles consulted and also the role of reviewers was specified.

Think-aloud protocol: After updating the selection criteria/process based on the review, two reviewers applied the selection criteria on five randomly chosen articles together. Reviewers used *"think-aloud protocol"* during selection. This helped to develop a common understanding of the criteria.

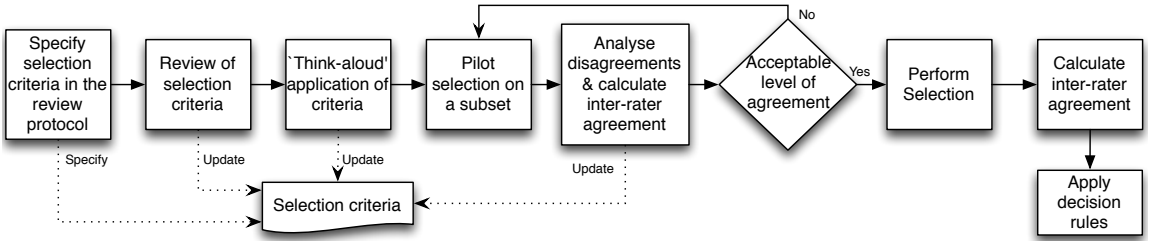

Figure 1: Overview of the selection process.

Pilot selection: After this step a pilot selection was performed where both reviewers independently applied the selection criteria on a randomly selected subset of 20 articles. Only two papers out of the 20 were classified differently by the two reviewers. These articles were discussed to understand why a difference existed and a consensus on interpretation of the criteria was reached and the protocol was updated. A high agreement upfront indicates that using "think-aloud protocol" helped improve the objectivity and understanding of the criteria.

Selection results: Two reviewers applied the selection criteria individually on the complete set of 1906 articles. The results are shown in Table 2. Inter-rater agreement was calculated to assure that the criteria were working well on the overall set of articles (percent agreement: 92.50 and Cohen's Kappa statistic: 0.73) which shows good agreement level.

Table 2: Results of applying the selection criteria

Cat. ID	# of articles	# of articles post-discussion	Remarks
A	96	106	Accepted for full-text reading
B	34	34	Accepted for full-text reading
C	122	174	Adaptive reading depth
D	30	0	Recategorised after discussion
E	82	0	Recategorised after discussion
F	1542	1592	Excluded

Out of the 1906 articles on which the selection criteria were applied, 112 articles ended up in categories *'D'* and *'E'*. These articles were discussed and recategorised in one of the other categories *'A'*, *'F'* or *'C'*. The final selection results after discussion are shown in the third column of Table 2.

Adaptive reading depth: A pilot application of the procedure for *"adaptive reading depth"* was done on a subset of five articles independently by both reviewers. The results were discussed to develop a shared interpretation. After this, we applied it on all 174 articles that were earlier categorized as *'C'*. 100 articles where both reviewers marked them as irrelevant were excluded and the remaining 74 articles were included for full-text reading (with percent agreement: 78.60 and Cohen's Kappa statistic: 0.53 showing moderate agreement). So in total we found 214 potential primary studies that were read in full-text. From an effort perspective, the adaptive reading took a maximum of 15 minutes per article.

Results of decision rules: From the 214 studies identified as potentially relevant articles, only 87 were included as primary studies. With this final list of primary studies, we can reflect on the decision choices during the selection process and loss/added value in terms of potentially relevant articles as shown in Table 3. Here ***"Overhead"*** is defined as the percentage of irrelevant articles that had to be analysed to identify relevant articles from a certain set of articles. ***"Contribution"*** is defined as the percentage of articles (from this set that became the primary studies in the review) of the total articles[1].

Table 3: Decisions categories and contributions.

Cat. ID	# of articles	Of which primary studies	Overhead	Contribution
A	96	51	46%	59%
B	34	11	67%	13%
C	122	16	86%	19%
D	30	3	90%	4%
E	82	6	92%	7%

Of the seven decision rules identified in existing secondary studies [5], decision rule D1:*''Majority vote i.e. a group of reviewers take a vote on the article and the decision of the majority is followed"* could not be evaluated in this study as only two reviewers applied the selection criteria. The rest of the six decisions, each can be seen as a combination of categories *'A'*-*'F'* from Table 1. These decision rules are listed below and mapped to our classification of disagreements in the first column of Table 4.

Table 4: Impact of strategies on selection results.

Strategy	# of articles	Of which primary studies	Overhead	Contribution
D3:{*A*}	96	51	46%	59%
D2:{*A*+*B*+*D*}	160	65	59%	75%
D6:{*A*+*B*+*C*}	252	78	69%	90%
D5:{*A*+*B*+*C*+*D*}	282	81	71%	94%
D4:{*A*+*B*+*C*+*E*}	334	84	74%	97%
D7:{*A*+*B*+*C*+*D*+*E*}	364	87	76%	100%

D2:*"At least one relevant then include"*
D3:*"All relevant then include"*
D4:*"At least one uncertain then include"*
D5:*"One irrelevant and one uncertain then exclude"*
D6:*"All researchers vote uncertain then include"*
D7:*"All irrelevant then exclude"*

These decision rules can be analysed for their effectiveness in identifying relevant primary studies as shown in Table 4. It can be seen that the existing rules vary from the most inclusive strategy (D7:{*A*+*B*+*C*+*D*+*E*}) to the least (D3:{*A*} which identifies less than 60% of primary studies). Thus, following *D3* and *D2* would have meant loss of 41% and 25% respectively, which could mean a significant threat to the reliability of the secondary study.

The most commonly used decision rule *D4* in existing secondary studies [5] is not the most inclusive strategy. This

[1]For *A* (using values in Table 3), overhead is (100*(96-51)/96) and contribution is 100*51/87 where 87 is the total number of primary studies.

strategy includes articles in category *'E'* which have a strong indication of being irrelevant and yet the articles in category *'D'* are excluded where one author has marked the article as relevant and other has marked it as irrelevant. Such articles could be just mistakes or may point out serious problems in the selection criteria used to guide the reviewers. Therefore, such articles should not be excluded without reflection.

5. LESSONS LEARNED

This section reports on the lessons learned from the use of the selection process in the systematic literature review.

Lesson 1: Assure repeatability by making strategies explicit: To ensure a repeatable selection process for secondary studies, the selection criteria, process and steps to resolve disagreements should be documented in a systematic way. Specifying decisions as described in this study will reduce ambiguity. For example, in the case for *D4* and *D6* it is unclear what should happen to papers in category *'B'* where some reviewers have considered the paper as relevant and others are uncertain.

Lesson 2: Use and interpret the inter-rater statistics with care: Large number of obviously unrelated articles found in automatic searches can skew the reviewers agreement levels (if measured by the percent agreement) and give false confidence in the objectivity of the criteria. Furthermore, often once the reviewers have an acceptable level of agreement they divide the remaining articles in disjoint sets to reviewers (cf. [1]). As shown by the results here a high agreement in the pilot does not mean that the overlap of articles between reviewers can be avoided. For example, in the given case, up to 20 articles (see contribution of *'B'*, *'D'* and *'E'* in Table 3) could have been lost if two reviewers were not involved.

Lesson 3: Level of disagreement indicates potential for finding relevant articles: Existing rules do not take this into consideration e.g. decision rule *D4* takes any article with a conflict further for full-text reading while *D3* excludes any articles where the reviewers have a difference of opinion. In this study, we took a unique approach to categorize the articles in categories demanding different levels of attention from the reviewers based on the type of disagreement (as shown in Table 1) and their potential for identifying primary studies. This indication was corroborated by the results of the systematic review as shown in Table 3, e.g. categories *'A'*, *'B'* and *'C'* where neither reviewer had considered the paper as irrelevant, contributed to identifying almost 90% of the primary studies. Similarly, articles in categories *'D'* and *'E'* where at least one author considered an article irrelevant only contributed in slightly over 10% of the primary studies.

Lesson 4: Following the most inclusive strategy leads to overhead with little gain: In the case of the systematic review under investigation we observed that the most inclusive strategy only provided little gains in terms of relevant articles identified with a considerable overhead, which is visible in Table 4.

With further evidence (by analysing these decision rules in secondary studies on other topics), we may conclude that there is a potential for investing much less time, and still achieving a good sample. Though, this will only be true if the sample on which the selection criteria are being applied is a true representative of the total population (c.f. [8]) and if there are certain consistent trends in results i.e. losing a few relevant articles will not alter the conclusions of the study altogether.

Lesson 5: Less inclusive strategies seem to lead to a loss of articles: The decision rules *D2* and *D3* result in significant loss of relevant studies as shown in Table 4. Unless we have more evidence to substantiate Lesson 4, we should avoid following such less inclusive strategies. Following the most inclusive strategy as a general principle, will reduce the likelihood of overlooking relevant articles and thus improve the consistency [8] of secondary studies.

Similarly, in the given case, there is a very small overhead of approximately 2% in moving from the most commonly used strategy *D4* to the most inclusive decision strategy *D7*. The extra articles that have to be reviewed in this strategy can be handled in a cost effective way by using adaptive reading depth [6]. As for articles with uncertainty and borderline cases the full-text reading is unnecessary.

Given the results, the combination of selection strategies used in this study is a good candidate-solution to become widely adopted for making study selection in secondary studies more repeatable. In the future, we would like to replicate this analysis on other secondary studies and deduce if general rules about the required level of inclusiveness can be discerned.

6. REFERENCES

[1] H. Edison, N. B. Ali, and R. Torkar. Towards innovation measurement in the software industry. *J. Syst. Software*, 86(5):1390–1407, 2013.

[2] K. R. Felizardo, G. F. Andery, F. V. Paulovich, R. Minghim, and J. C. Maldonado. A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Inform. Software Tech.*, 54(10):1079–1091, 2012.

[3] B. Kitchenham. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Department of Computer Science, Keele University, ST5 5BG, UK, 2007.

[4] B. Kitchenham and P. Brereton. A systematic review of systematic review process research in software engineering. *Inform. Software Tech.*, 55(12):2049–2075, 2013.

[5] K. Petersen and N. B. Ali. Identifying strategies for study selection in systematic reviews and maps. In *Proc. of the 5th Int. Symp. on Empirical Software Engineering and Measurement, ESEM 2011*, pages 351–354, 2011.

[6] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson. Systematic mapping studies in software engineering. In *Proc. of the 12th Int. Conf. on Evaluation & Assessment in Software Engineering (EASE 2008)*, pages 71–80, 2008.

[7] Y. Sun, Y. Yang, H. Zhang, W. Zhang, and Q. Wang. Towards evidence-based ontology for supporting systematic literature review. In *Proc. of the 16th Int. Conference on Evaluation & Assessment in Software Engineering (EASE 2012)*, pages 171–175. IET, 2012.

[8] C. Wohlin, P. Runeson, P. A. da Mota Silveira Neto, E. Engström, I. do Carmo Machado, and E. S. de Almeida. On the reliability of mapping studies in software engineering. *J. Syst. Software*, 86(10):2594–2610, 2013.