Successful Combination of Database Search and Snowballing for Identification of Primary Studies in Systematic Literature Studies

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

[a] *Blekinge Institute of Technology, Karlskrona, Sweden*
[b] *Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil*
[c] *Federal Technological University of Paraná, Cornélio Procópio, Brazil*

Abstract

Background: A good search strategy is essential for a successful systematic literature study. Historically, database searches have been the norm, which was later complemented with snowball searches. Our conjecture is that we can perform even better searches if combining these two search approaches, referred to as a hybrid search strategy.
Objective: Our main objective was to compare and evaluate a hybrid search strategy. Furthermore, we compared four alternative hybrid search strategies to assess whether we could identify more cost–efficient ways of searching for relevant primary studies.
Method: To compare and evaluate the hybrid search strategy, we replicated the search procedure in a systematic literature review (SLR) on industry–academia collaboration in software engineering. The SLR used a more "traditional" approach to searching for relevant articles for an SLR, while our replication was executed using a hybrid search strategy.
Results: In our evaluation, the hybrid search strategy was superior in identifying relevant primary studies. It identified 30% more primary studies and even more studies when focusing only on peer–reviewed articles. To embrace individual viewpoints when assessing research articles and minimise the risk

[*]Corresponding author
Email addresses: claes.wohlin@bth.se (Claes Wohlin [*,]), kalinowski@inf.puc-rio.br (Marcos Kalinowski), katiascannavino@utfpr.edu.br (Katia Romero Felizardo), emilia.mendes@bth.se (Emilia Mendes)

of missing primary studies, we introduced two new concepts, *wild cards* and *borderline articles*, when performing systematic literature studies.
Conclusions: The hybrid search strategy is a strong contender for being used when performing systematic literature studies. Furthermore, alternative hybrid search strategies may be viable if selected wisely in relation to the start set for snowballing. Finally, the two new concepts were judged as essential to cater for different individual judgements and to minimise the risk of excluding primary studies that ought to be included.

Keywords:
Systematic Literature Reviews, Hybrid Search, Snowballing, Scopus

---

1. Introduction

According to the guidelines for performing systematic literature studies authored by Kitchenham and Charters (2007), a secondary study is defined as "A study that reviews all the primary studies relating to a specific research question with the aim of integrating/synthesising evidence related to a specific research question." We use the term systematic literature study (SLS) as a generic term, including both systematic literature reviews and systematic mapping studies.

Given that the definition highlights that we are expected to find "all" the primary studies, the search strategy becomes essential in achieving (or at least moving towards) this goal. The two main approaches to search for primary studies are database search and snowballing. The Kitchenham and Charters (2007) guidelines describe database search, and Wohlin (2014) provides guidelines for using snowballing. These two search strategies are compared by Jalali and Wohlin (2012). However, an alternative is to combine them. The concept of a hybrid approach was proposed by Wohlin (2014) and listed as an area for future research. Since then, Mourão et al. (2017) investigated one hybrid search strategy, which was later extended by Mourão et al. (2020) to four alternative hybrid search strategies that were also evaluated.

A hybrid search strategy is *defined* herein as follows: A hybrid search strategy is the pre-planned integration of at least two systematic approaches to searching for articles for an SLS. For example, performing systematic searches in one or more digital databases or indexing services and then snowballing from all the relevant articles found in the previous searches. A hybrid

search strategy should not be confused with performing several complementary searches, for example, a database search complemented with searching in some specific conference proceedings. Furthermore, performing a database search and then snowballing from a sample of the articles found in that search is also not a hybrid search strategy according to our definition.

At the outset of the research, the overall objective was to compare and evaluate the search strategy employed in a previously published SLS with a *replication* using a hybrid search strategy, and to compare the four different hybrid search strategies presented by Mourão et al. (2020).

Concerning replication as a research method, according to Porte (2013), replication is a key method for acceptance of new knowledge. Shull et al. (2008) discuss the role and importance of replication in Software Engineering (SE). Replications are scarce in SE since they are labour intensive, and their scientific value may be challenged in terms of novelty. In relation to SLSs, replications exist, for example, some researchers have relied on intentional replication (MacDonell et al., 2010), and other researchers took the opportunity to compare SLRs when two SLRs on the same topic were published (Wohlin et al., 2013). The replication herein includes a different procedure (search strategy) and different researchers, and it may be classified as being a conceptual replication in SE (Shull et al., 2008).

Based on the above objective, we posed the following two research questions:

- RQ1: How many articles does the replication using a hybrid search strategy find in relation to the original SLS?
- RQ2: How do the four alternative hybrid search strategies compare to each other concerning the number of articles identified?

The four hybrid search strategies are presented in Section 3.1. However, note that, as the research progressed, some other findings emerged based on the overall objective.

In relation to the research questions, the following contributions are made:

- A replication of the search procedure in an SLS on industry–academia collaboration is presented; the study is reported as an SLR. The replication includes approximately 30% more relevant articles than the original SLR, and hence it provides an added value to the original SLR on industry–academia collaboration (IAC). A comparison between the

original SLR and the replication is provided to discuss the differences and similarities.

- The four hybrid strategies are compared, and it is concluded that the full-fledged hybrid search strategy is the best, although it comes at the cost of more work.

In addition to the overall objective, the article contributes with the following:

- Two new concepts are introduced to mitigate the different judgements of reviewers performing an SLS. The first concept is so-called *wild cards* to allow for taking an article to full–text assessment even if it does not meet the formal inclusion criteria used to filter the studies that are moved to the next stage. Secondly, we introduce the concept of *borderline articles*, which are articles that were close to being included in an SLS. Thus, we suggest that articles being close to inclusion should be listed separately, and not in a general listing of excluded articles. The objective of keeping track of them is to allow for others to make their own judgement concerning these articles.

- Based on the outcome, in this particular case, a new hybrid search strategy is put forward for further research. We have chosen to call it an adaptive hybrid search strategy.

Here, the hybrid search strategy is implemented using a search string in Scopus to identify article candidates for inclusion. These articles are assessed, and for those included, both backward and forward snowballing are performed using the guidelines for snowballing (Wohlin, 2014). In addition, to evaluate more cost-efficient hybrid methods, three snowballing alternatives are considered. Mourão et al. (2020) introduced the four alternatives. They are summarised in Section 3.1.

The main advantage of a hybrid search strategy is that it is intended to mitigate some of the drawbacks of using different implementations of databases or indexing services and ensures that we have a good start set for snowballing.

The remainder of this article is organised as follows. In Section 2, we introduce related work, and the research design is presented in Section 3. Section 4 details the preparations for the evaluation, followed by the execution of the replication in Section 5. Next, Section 6 presents the results in

the form of an evaluation, i.e., a comparison between the original SLR and the replication, and threats to validity are discussed in Section 7. Finally, Section 8 concludes the work concerning the evaluation of the hybrid search strategy and provides suggestions on future work.

2. Related work

This section chronologically summarises the state–of–the–art of search strategies for SLSs in SE. As demonstrated in Figure 1 and Table 1, up to now, database search (DBS) and snowballing (SB) are the leading approaches adopted in SE to search for primary studies to include in SLSs.

Table 1: Chronological view of search strategies for SLSs in SE.

Comparison	**Hybrid search?**	**Reference**
Searches exercising terms		Dieste et al. (2009)
Reference-based search		Skoglund and Runeson (2009)
MS with DBS		Kitchenham et al. (2010)
MS with DBS		Zhang et al. (2011)
DBS with Google Scholar + BS		Jalali and Wohlin (2012)
Google Scholar + BS*FS with DBS		Wohlin (2014)
Google Scholar + BS*FS with DBS		Badampudi et al. (2015)
FS with DBS		Felizardo et al. (2016)
FS with DBS		Wohlin (2016)
DBS with Scopus + BS; Scopus + FS; Scopus + BS||FS	✓	Mourão et al. (2017)
FS with DBS		Felizardo et al. (2018)
FS with DBS		Mendes et al. (2019)
DBS with Scopus + BS*FS; Scopus + BS||FS; Scopus + BS+FS; Scopus + FS+BS	✓	Mourão et al. (2020)
Legend: Manual Search (MS); Database Search (DBS) and Backward/Forward Snowballing (BS/FS)		

Dieste et al. (2009) analysed the effects of using different terms and combinations of terms to find an optimum search strategy for use in SLSs of SE experiments. In total, 29 search strategies were investigated, and they concluded that optimising search strings for retrieving relevant SE experiments is not a straightforward task.

Skoglund and Runeson (2009) proposed and evaluated a search strategy that uses semantic information in references between articles to find rele-

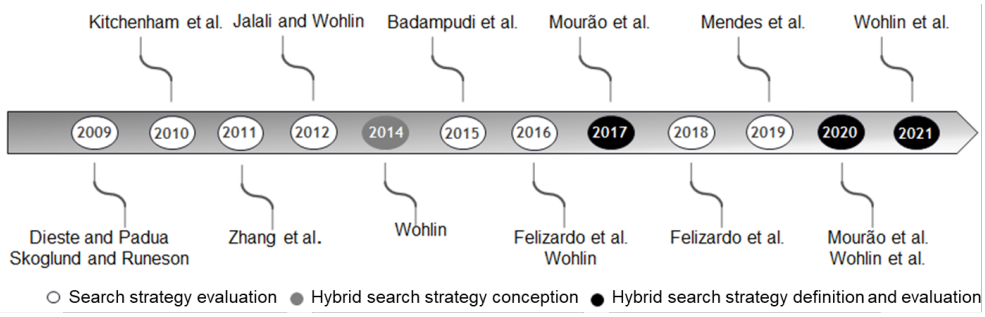

Figure 1: Search strategies for SLRs in SE shown in chronological order.

vant articles. The strategy is composed of four steps: (1) identification of a "take-off article" – a relevant article on the SLS topic, which is systematically chosen; (2) candidate articles referenced by the "take-off article" – the reference list of the"take-off article" is analysed to reveal other relevant articles and the reference list of each revealed article is also analysed; (3) identification of "cardinal articles" – those articles referenced more than others; and (4) articles from external sources referencing the "cardinal articles" – these external articles may be identified using digital libraries. They evaluated their strategy over three SLRs and observed that the results were not satisfactory for two of the SLRs in terms of precision and relative recall.

Kitchenham et al. (2010) undertook two participant-observer case studies aiming to compare both the use of manual searches with broader searches and the quality of the articles retrieved. The original SLR replicated in their study relates to SLRs in SE, and the manual search covered selected journals and conferences. In the replicated study, searches including digital libraries and general indexing services were carried out. Their results indicated that searches in digital libraries and indexing services found more articles than manual searches. Still, broader searches may require more time and effort, and the articles found might be of low quality, as highlighted by Kitchenham et al. (2010).

Zhang et al. (2011) proposed a search strategy for retrieving articles for SLSs in SE. The strategy is based on the concept of "*quasi-gold standard*" (QGS) and integrates manual and database searches. Initially, a pool of articles is manually identified. Next, this set of articles is used to elicit relevant terms to elaborate a search string for database search and validate the

Table 2: Main results of search strategies comparison.

Comparison	**Main result**	**Reference**
Different terms	It is not straightforward to optimise a search strategy for retrieving SE experiments articles.	Dieste et al. (2009).
Reference list	The reference-based strategy seems to work better in specific SE topics.	Skoglund and Runeson (2009)
MS and DBS	DBS searches were able to find more articles than MS, but potentially with lower quality.	Kitchenham et al. (2010).
MS + DBS	A QGS-search strategy.	Zhang et al. (2011)
DBS and BS	Similar results for both search strategies.	Jalali and Wohlin (2012).
SB and DBS	SB may be a potential alternative to DB searches. Efficiencies of SB and DB searches are equivalent.	Badampudi et al. (2015). Wohlin (2014)
FS and DBS	DBS and FS are comparable in finding articles.	Felizardo et al. (2016, 2018); Mendes et al. (2019) and Wohlin (2016)
DBS with hybrid search strategies	Hybrid strategies may be a contender for SLRs.	Mourão et al. (2017) and Mourão et al. (2020)
Manual Search (MS); Database Search (DBS); Quasi-Gold Standard (QGS) Snowballing (SB) = Backward/Forward Snowballing (BS/FS)		

search performance (quasi-sensitivity). Finally, the performance is calculated by the number of relevant articles retrieved from the manual search through the database search, divided by the pool size of QGS. On the one hand, the search process terminates if the quasi-sensitivity is $\geqslant 80\%$ (acceptable performance), and the results from the database search, and the QGS are merged. On the other hand, the search string is refined until the performance becomes acceptable. Two participant-observer case studies were performed to demonstrate and evaluate the QGS–search strategy, which was considered to improve the rigour of search processes in SLSs.

Jalali and Wohlin (2012) compared two SLRs on Agile practices in GSE using DBS and Backward Snowballing (BS), respectively. The same authors performed the SLRs for the same time interval to enable comparison. The start set of articles for the BS approach was generated through a search in Google Scholar, and then BS was applied based on the articles found. Both studies kept the search terms and keywords as similar as possible. They noted

that regardless of strategy, most of the included articles were the same, and both studies' conclusions were quite similar; therefore they concluded that the SLRs' results were not dependent on their search strategy. They also learned that SB might be more efficient when the keywords for searching include general terms, since it reduces DBS noise.

Wohlin (2014) established guidelines for performing snowballing (SB) as a search strategy for SLSs. The strategy defines running Backward Snowballing (BS) and Forward Snowballing (FS) in iterations. The application of SB was also compared with an SLR that initially used DBS as a search strategy. The results indicated that SB could replace the search in several different databases, hence being an alternative search strategy to use when performing SLSs. Furthermore, Wohlin (2014) puts forward the idea of a hybrid search strategy for further research, see Figure 1.

The study of Badampudi et al. (2015) complements the studies by Jalali and Wohlin (2012) and Wohlin (2014). It evaluated the effectiveness (number of articles included concerning the total number of articles reviewed) and the reliability (the ability to capture all relevant articles) of using SB (BS and FS) as a search strategy in SLSs, when compared with DBS. In their study, different researchers performed DBS and SB independently, in the same period. Like Jalali and Wohlin (2012), the SB search was carried out by first creating a start set using Google Scholar. Next, BS and FS were applied iteratively based on the articles found. The outcomes of the study by Badampudi et al. (2015) are similar to those reported by Jalali and Wohlin (2012) and Wohlin (2014); they found that the effectiveness of SB is comparable to DBS and that its reliability is highly dependent on the creation of a suitable start set.

Between 2016 and 2020, Felizardo et al. (2016), Wohlin (2016) and Mendes et al. (2019) have investigated SB as a search strategy for updating SLRs. They used FS to replicate a second-generation study performed using double DB searches; i.e., DB searches were performed twice covering different periods, i.e., for the original SLR and its update. The key conclusion is that FS can find relevant articles updating SLRs in SE. Felizardo et al. (2018) evaluated the use of specific and generic (e.g., Scopus or Google Scholar) databases/services for applying FS to update SLRs. They concluded that using an indexing service (Google Scholar) is sufficient to find articles. These studies together enabled the definition of guidelines focused on recommendations for a search strategy specific for updating SLRs Wohlin et al. (2020). The main recommendations include employing FS using Google Scholar and using the SLR and its primary studies to compose the starting set of articles

for the FS.

As shown in Table 2, only two studies, Mourão et al. (2017) and Mourão et al. (2020), have addressed the use of hybrid search strategies in the SE area. In 2017, Mourão et al. (2017) proposed and evaluated one hybrid search strategy. More recently, in 2020, the research was extended to four hybrid search strategies that combine DBS and SB in different ways. These four strategies are described in detail in Section 3. The authors compared the outcome from DB searches, snowballing, and hybrid strategies and concluded that using a hybrid search strategy involving a representative digital library (e.g., Scopus) and parallel or sequential snowballing may be an appropriate alternative for searching for candidates articles in SLSs.

Unlike the studies previously mentioned, which directly contrasted DBS with SB, our study evaluates hybrid strategies. The research reported in this article builds upon the two previously mentioned studies by Mourão et al. (2017) and Mourão et al. (2020), where the hybrid search strategy performed well. However, in particular, we evaluate the four hybrid search strategies presented in Mourão et al. (2020) over one existing SLR in SE. The SLR adopted a database search strategy, followed by complementary searches using BS and FS, and further complementary searches in specific venues to expand the candidate articles. Finally, we also suggest a new hybrid search strategy for further investigation.

3. Research design

The research team for evaluating the hybrid search strategies has previously conducted research on SLRs. It includes updating an existing SLR, which is closely related to replicating an SLR. Thus, the team has collective experience in this type of research.

3.1. Four hybrid search strategies

The four hybrid search strategies all start with a search in Scopus to create a start set. Next, we have the following four alternatives for hybrid search strategies, as described in Mourão et al. (2020) and illustrated in Figure 2:

1. **Scopus followed by full BS and FS**: New articles are identified through backward and forward snowballing based on the start set. It is done for all articles meeting the inclusion criteria. This alternative is the full-fledged search strategy according to the guidelines for snowballing in Wohlin (2014).

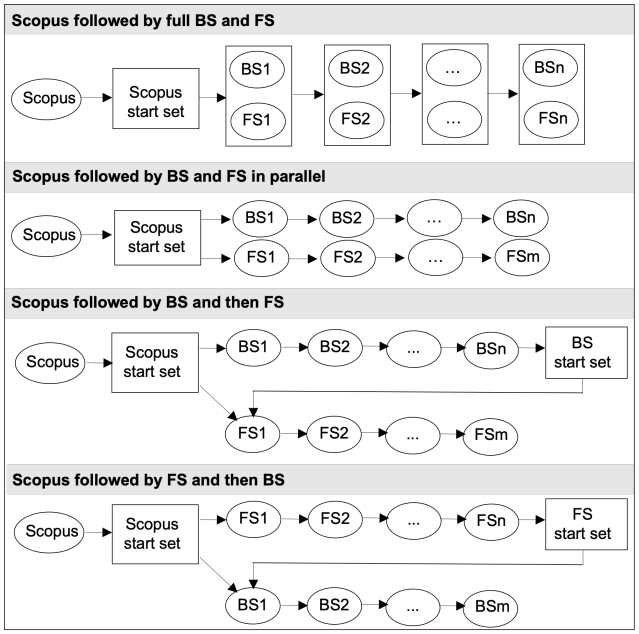

Figure 2: Four alternative hybrid search strategies.

2. **Scopus followed by BS and FS in parallel**: In this alternative, backward and forward snowballing run as two separate processes using the same start set. In other words, the articles obtained by backward snowballing are not subject to forward snowballing and vice-versa. This strategy was first introduced by Mourão et al. (2017) to increase the precision without compromising the recall.
3. **Scopus followed by BS and then FS**: Here, backward snowballing is performed using the start set following the guidelines for snowballing in Wohlin (2014). After finishing all backward snowballing iterations, forward snowballing begins by using the start set from Scopus and the articles found through the backward snowballing. Backward snowballing is not done for the articles found through forward snowballing.
4. **Scopus followed by FS and then BS**: The fourth alternative is similar to the previous option, although starting with forward snowballing. Thus, forward snowballing is performed using the start set, as per the guidelines for snowballing in Wohlin (2014). After finishing all forward snowballing iterations, backward snowballing begins, using

the start set from Scopus and the articles found through the forward snowballing. Forward snowballing is not done for the articles found through backward snowballing.

The objective is to use the first hybrid search strategy and then to carefully track when different articles are found to "simulate" the other three hybrid search strategies.

3.2. Research approach

To address the research questions, we defined a research approach inspired by the general process for problem–solving, as formulated by Agnew and Pyke (2007), which has three steps: observe–think–test. The three steps are reformulated herein as prepare–execute–evaluate to align with our research focus. Thus, the research approach is as follows:

Prepare: A search was performed in Google Scholar to identify candidate SLRs, using the following search string: "systematic literature review software engineering". The search used the time interval 2015-2019, and patents and citations were not ticked. The criteria and process for selecting an SLR for replication are presented in Appendix A. It resulted in selecting an SLR investigating industry–academia collaboration in SE, authored by Garousi et al. (2016).

Once the SLR was decided upon, it was distributed among the authors. It was agreed that we should only look at the sections preceding the presentation of the results. Consequently, it was impossible to use SLR quality as one selection criterion. Reading part of the original SLR was needed to understand the research design and to replicate the selection process, in particular when applying the inclusion and exclusion criteria. The first author prepared Excel sheets to support the assessment, which mitigated some of the challenges due to the authors being distributed.

Execute: A start set for snowballing (both backward and forward snowballing) was identified for the topic of the selected SLR. The start set was determined by searching in Scopus and evaluating the resulting articles. Scopus was chosen motivated by the results found in Mourão et al. (2020). The inclusion of articles in our replicated SLR was based on the criteria documented in the published original SLR. Once a start set was identified, backward and forward snowballing were performed to determine further articles to include, using the guidelines in Wohlin (2014).

Evaluate: An analysis was carried out to address the research questions. The comparison was made concerning the included articles and the SLR's topic coverage. Based on the analysis, the observations from the evaluation are reported, and the different search strategies are contrasted.

The three steps in the evaluation process are described in more detail in the following three sections.

4. Prepare

The preparations for the replication include some activities aimed to have a common understanding of both the original SLR and how to execute the replication. Section 4.1 describes how the search was performed by Garousi et al. (2016). It summarises their search, inclusion/exclusion criteria and inclusion/exclusion process. Section 4.2 presents our preparation for carrying out the replication, including introducing the concepts of wild cards and borderline articles. Moreover, it describes how the supporting Excel sheets were formulated, and how the four authors divided the review work. Finally, it presents how the inclusion/exclusion process from Garousi et al. (2016) was implemented for the replication.

4.1. Essential aspects of the selected SLR

The objective of the selected SLR by Garousi et al. (2016) is stated as follows: "To identify (a) the challenges to avoid risks to the collaboration by being aware of the challenges, (b) the best practices to provide an inventory of practices (patterns) allowing for an informed choice of practices to use when planning and conducting collaborative projects." Thus, articles describing collaboration per se between academia and industry should not be included since the focus is on challenges and best practices in the collaboration. Furthermore, Garousi et al. (2016) described the context of their SLR as being experiences and lessons learnt in industry–academia collaboration, as reported either by researchers or practitioners. The focus of the chosen SLR is summarised in three research questions by Garousi et al. (2016). They look at collaboration models between industry and academia, challenges and impediments highlighted, and finally, patterns in terms of best practices. To identify articles, Garousi et al. (2016) used the following search process:

- They searched in three different sources: IEEE Xplore, ACM Digital Library and Google Scholar.

- Fifty-four (3*2*3*3) search strings were used by combing four different listings and searching for all combinations. The listings can be found in Table 3. As an example of the 54 search strings, we have "Industry AND Academia AND Collaboration AND 'Software engineering'." Thus, Garousi et al. (2016) took all combinations of the four listings in Table 3.

- The searches were performed respectively in January and February 2015. Only the articles available at this time were included in the pool of potential studies for inclusion in the SLR.

- To complement the searches in IEEE Xplore, ACM Digital Library and Google Scholar, and to ensure that they did not miss any relevant articles, Garousi et al. (2016) randomly selected five articles from the searches and performed backward and forward snowballing (Wohlin, 2014). It is not reported how the randomisation was done or which five articles were included in the snowballing.

- Finally, Garousi et al. (2016) also looked at some specific venues for additional articles. The selected venues are not reported in their article.

Listing 1	**Listing 2**	**Listing 3**	**Listing 4**
Industry	Academia	Collaboration	Software engineering
Practice	Theory	Relationship	Software
University		Relation	IT

Table 3: List of keywords for creating search strings.

The search strategy is essential since our hybrid approach was to be compared to the original SLR's search strategy.

When it comes to the article's inclusion/exclusion criteria, the following is stated in Garousi et al. (2016):

- The main criterion concerns whether a given study presents relevant findings for industry and academia collaboration in SE.

- Only articles written in English are included.

- Only articles available electronically are included.

- If a conference article has a more recent journal version, then the journal article is included, and the conference article is excluded.

- Only the most recent article is included if multiple studies with the same title by the same authors are found.

The inclusion/exclusion process is described as follows:

- All three authors of Garousi et al. (2016) looked at the pool of potential articles to be included. It is unclear how the initial pool was identified based on the search process, i.e., the initial screening of articles is not described.
- For each article, Garousi et al. (2016) assigned individual scores: 0 – exclude, 1 – uncertain, and 2 – include.
- The initial scores were based on reviewing each article's title, abstract, and keywords. If this was insufficient to assign a score, the authors of Garousi et al. (2016) looked at the full article. However, it is unclear, when they looked at the full articles, whether each individual decided separately, or whether it was a joint decision.
- Articles were tentatively included as candidates whenever the total score was four or higher, and excluded otherwise. Final inclusion/exclusion was based on the three authors independently assigning scores to the full articles.

Garousi et al. (2016) do not report any agreement calculations, for example, using Fleiss kappa. Furthermore, they do not discuss performing any quality assessment of the primary studies, and hence it is assumed here that they included all the primary studies identified.

4.2. Our preparations

4.2.1. Introducing two novel concepts

Two novel concepts were introduced in the reviewing process to address researchers' different judgements while assessing the articles. The two new concepts are:

- Wild cards – A *wild card* is, in our context, an article that fails to qualify in the usual way, i.e., by fulfilling the score needed to be included. The concept is taken from sports, where certain persons may be invited to participate in a tournament even if not meeting the qualification criteria. Each reviewer is allowed to nominate a wild card as described in Section 4.2.4.

- Borderline articles – A *borderline article* is defined as an article receiving a score of three in the full–text assessment. To achieve such score, either all three reviewers are uncertain, or one reviewer wants to include the article, and one reviewer is uncertain. Thus, the articles being assessed are placed in three categories: "include", "exclude", and "exclude, but borderline". Articles in the third category are recorded separately. The objective is to provide transparency into which articles were close to being included. Furthermore, it allows readers of our review to assess the borderline articles themselves and decide whether they judge them as either relevant or not.

We believe that both concepts embrace differences in opinions, and in doing so, provide much more flexibility than, for example, having two persons convincing a third person that a paper should be excluded.

4.2.2. Supporting Excel sheets

The first author developed Excel sheets to support the start set identification and to perform both backward and forward snowballing. The Excel sheets included links to the articles identified to help the individuals judging these articles.

4.2.3. Reviewers of articles

The first three authors acted as reviewers of the articles found through the Scopus' search and took part in the backward and forward snowballing. It was essential to involve three researchers to mimic the process used by the the original SLR's authors of. The fourth author acted as a backup in the reviewing process.

4.2.4. Inclusion/exclusion process

We used the same inclusion/exclusion process as in the original SLR, see Section 4.1. All articles found were assigned scores, as also done in the original SLR, i.e. first on title, abstract and keywords, and then for tentative candidates, the full article was assessed. Note that if a reviewer perceives that it is infeasible to give a score without looking at the full article, the score assigned is one, for "uncertain".

Moreover, the inclusion/exclusion process in the original SLR was adapted as follows. Each reviewer was allowed to nominate a "wild card" if it was perceived that the full article should be assessed, even if the article did not meet the score needed to be included for full–text assessment.

If an article was nominated as a wild card by more than one reviewer, each of these reviewers was allowed to nominate an additional article. This was repeated until the number of included wild card articles was at least the same as the number of reviewers.

Wild cards were not used in the full–text review, i.e., only articles meeting the scoring threshold were included as a primary study in the SLR. However, articles receiving a score of three were recorded separately in a list of "borderline articles". As in the original SLR, quality assessment was not performed for the primary studies, because it is not critical when identifying primary studies. However, it would be essential when synthesising the evidence from the selected primary studies. The latter is out of the replication's scope, given the focus on the search strategy.

5. Execute

5.1. Search in Scopus

The first step in the hybrid search strategy is to identify a start set for snowballing. As stated in the research design, Scopus was chosen to search for articles to form the start set.

The search in Scopus was limited to the five years preceding the date when the selected SLR was performed, i.e., not published. It was decided that the start set should include at least five articles, but preferably ten articles. This was done because having only a few articles may bias the snowballing and result in missing relevant articles. Thus, a sufficiently large subset is needed. However, it is hard to determine an optimal size since it depends on the number of articles (which is unknown upfront) and the number of active researchers in the area (since authors tend to cite their own relevant articles). The start set should contain articles to be included among the primary studies in the SLR. If less than five articles were found, the objective was to increase the time interval by one year at a time until the start set had at least five articles.

The candidate articles for the start set were found using the following process:

1. Given the area of the selected SLR to replicate, the following search string was formulated to be used in Scopus: industry AND academia AND collaboration AND software AND engineering.

2. Given that the selected SLR was performed in early 2015, although published in 2016, the first search was done for the time interval 2010-2014.
3. Only articles published in either journals or conferences were included, which means that all the included articles were peer–reviewed. Articles of types: book, book chapter, conference review and notes are excluded from the Scopus search.

In total 40 articles were found using Scopus. Their titles were included in an Excel sheet with links to the articles, based on Scopus's listing. This ensured that the authors used the same information when providing scores for the different articles.

5.2. Identification of start set

The articles resulting from the search were evaluated based on the the original SLR's inclusion and exclusion criteria, and rated as described in Section 4.2.4. Given that snowballing should only be performed on the articles to be included in the final set, it was necessary to look at the full articles identified in the scoring procedure, including the wild cards, before performing the next step in snowballing. However, it was unclear whether articles relating to the collaboration between industry and academia in education were included or excluded in the original SLR. To obtain a clarification, one of the original SLR's authors was contacted, and informed that the focus was on research collaborations. Thus, given that the objective was replication, we also focused on articles concerning collaboration in research.

In total, 15 articles fulfilled our inclusion criteria and went into full–text reading. They included twelve articles selected based on their scores and three articles as wild cards. The 15 articles were assessed by the three reviewers based on scores.

This resulted in nine articles being included. All nine had each a total score of six (each reviewer gave a score of two, see Section 4.1), meaning that all three reviewers wanted to include the articles. Thus, there was a consensus for inclusion among the three reviewers. Six of the original 15 articles were excluded, and the three reviewers' opinions about these six articles varied. For five of these articles, one reviewer wanted to include the article. However, it should be noted that only one article had a score of three, and hence only one article was borderline to be included. It is also worth mentioning that none of the three wild cards nominated was included.

Thus, the assessment on title, abstract and keywords was well aligned with the assessment on full–text reading.

The different views among the three reviewers were not critical for the main objective in the article, i.e., to assess and compare the various search strategies applied in the original SLR and the replication presented here. Thus, even if an article was excluded, it is documented that it was found. This was considered when comparing the search strategies used in the original SLR and the replication.

Although only journal and conference articles were considered when identifying the start set, book chapters and workshop articles were also included when performing snowballing. We did not ask the original SLR's authors about this aspect, since we wanted to be as independent as possible when performing the search. The only exception occurred when we asked about industry–academia collaboration in education, where the answer was a simple yes/no. If we asked about publication types included, this could lead to a discussion regarding publication types, which we wanted to avoid. Furthermore, it is substantially easier to remove them later on, than to add them. Book chapters and workshop articles were included under the assumption that they were peer–reviewed. We did not consider other types of publications such as, for example, books and theses, since they have most likely not been through the same peer–review as research articles, including book chapters.

5.3. Backward snowballing

In backward snowballing, the following process was used:

1. An Excel sheet was created, containing a listing of the articles included based on the Scopus search. These articles formed the start set for the first round of backward snowballing. In the second round, articles included from the previous round, i.e., first round, were included, and so forth.
2. Each reviewer was asked to go through the reference lists for the included articles. Articles in these reference lists judged to definitively be out of the scope of the systematic literature review were not moved to the Excel sheet. The judgement was based on the article's title and how and where it was cited. Articles that could potentially be of interest were included by each reviewer into an individual Excel sheet.

3. Each reviewer assessed the articles put into their respective Excel sheet based on title, abstract and keywords using the same procedure as when reviewing the articles found through the Scopus search, as described in Section 4.2.4.
4. The assessment based on title, abstract and keywords was coordinated. It means that if one of the reviewers gave a score of 1 or 2 for an article not assessed by one or two of the other reviewers, it also had to be assessed by those who have not assessed it. Thus, all articles with scores of 1 or 2 from at least one reviewer were also assessed by all reviewers.
5. The full–text assessment was performed as described in Section 4.2.4.

It should be observed that we were careful to track all included articles in terms of when they were found. There was a need to track when an article was identified for inclusion for the first time. However, to evaluate all four hybrid search strategies, we needed to keep track of all the instances when an article was found.

5.4. Forward snowballing

In forward snowballing, the following process was used:

1. The article's title was pasted into Google Scholar's search box to find it. For each title, the citations were identified also using Google Scholar. The search for citations was limited to articles published in 2014 or earlier. Patents and citations were unticked, i.e., not considered in the search.
2. Links to the peer–reviewed articles citing each article were added to an Excel sheet by the first author to simplify the assessment.
3. The articles were assessed as described in Section 4.2.4.

5.5. Article identification

Through the hybrid search strategy, we identified in total 1942 publications of interest. The publications came from the start set, backward and forward snowballing, respectively, as follows: 40 publications from Scopus (start set), 839 publications from Backward Snowballing (BS), and 1063 publications from Forward Snowballing (FS). The inclusion/exclusion process is illustrated in Figure 3. It includes four steps:

- Step 1 – Screening
  In total 390 articles were removed through screening, i.e., without assessment. The main reasons for the removal was that the identified publications did not meet the criteria of being peer-reviewed, or they were duplicates from articles already assessed.

- Step 2 – Title, abstract and keyword level assessment
  The three reviewers looked at all 1402 potential articles assessing each article's title, abstract, and keywords. It resulted in 60 articles having scores four and higher (sum 4 – 16, sum 5 – 14; sum 6 – 30). These articles were moved to full text assessment.

- Step 3 – Wild cards
  During Step 3, the reviewers could nominate wild cards. In total 18 wild cards were nominated from articles receiving a score of 2 or 3 in the assessment in Step 2 as shown in Figure 3.

- Step 4 – Full text level assessment
  Finally, in total 78 articles went into full text assessment. It resulted in the inclusion of 43 articles, and ten articles were listed as borderline. The remaining 25 articles were excluded after reading their full text.

Furthermore, the Fleiss kappa value was calculated to evaluate the level of agreement between the three reviewers. On the abstract level, the Fleiss kappa becomes 0.68, which is a good agreement. On the one hand, the kappa value is affected by the agreement concerning articles that do not contribute to the area of industry–academia collaboration. On the other hand, the wild cards help mitigate some of the reviewers' disagreements. On the full–text level, the Fleiss kappa is 0.43, which is a moderate agreement. The lower kappa value is due to the reviewers having different opinions concerning what constitutes a sufficient contribution about industry–academia collaboration. The main disagreement concerns the articles with a summed score of three. Given the different views, we have listed these articles separately as borderline articles, allowing readers to judge the ten borderline articles.

The articles identified in the replication should be contrasted with the 33 publications listed in the original SLR. Figure 4 illustrates the articles identified based on the nine articles of the start set throughout the snowballing iterations. The articles included in the start set and after each backward and forward snowballing iteration are listed in Appendix B, which also lists the borderline articles. The results are further elaborated in Section 6.

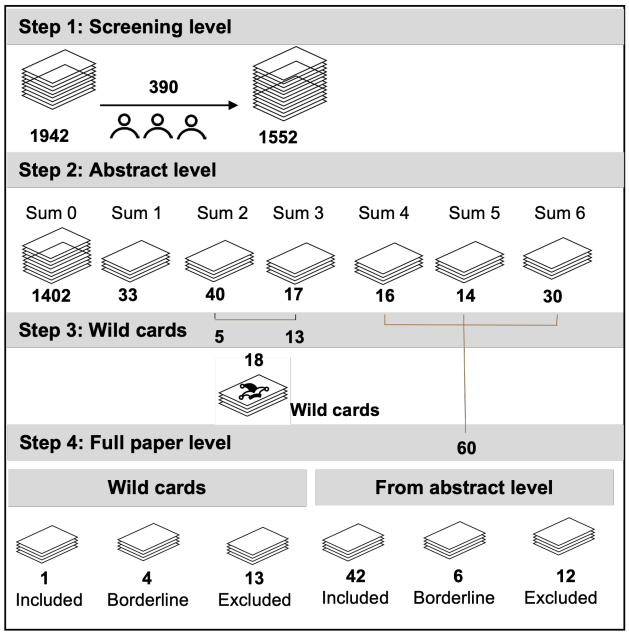

Figure 3: Selecting articles using the concepts of wild cards and borderline articles.

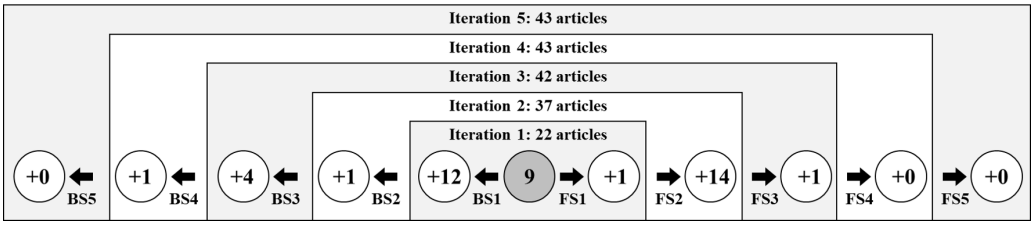

Figure 4: Articles identified throughout the snowballing iterations based on the start set.

6. Evaluate

The results include a discussion on the two new concepts introduced in Section 6.1. Four essential aspects related to the evaluation are provided in Section 6.2. In Section 6.3, a comparison of the original SLR and its replication is given in terms of the articles identified through the two different search strategies. In Section 6.4, the different hybrid search strategies are compared. Finally, some reflections concerning the execution of the replication are presented in Section 6.5.

6.1. The two new concepts

Both new concepts, i.e., wild card and borderline article, were helpful. One of the wild card articles was included after the full–text assessment. The article by Punter and van de Laar (2010) was nominated as a wild card and then included after full–text assessment. This article was not included in the original SLR. Furthermore, four nominated wild cards turned into borderline articles. These four articles are identified as 70, 73, 115, and 140 in Appendix B. Article 140 is included in the original SLR, while the other three articles are not.

The wild card concept resulted in one article being included that would otherwise have been excluded. Furthermore, it helped identify four more borderline articles. Moreover, such a concept is also helpful since it illustrates which articles were close to being included. In this way, readers may decide for themselves whether the articles contain valuable findings. In summary, we argue that both concepts are useful additions to the current standard practice of performing SLRs. Wild cards help mitigate different opinions among the reviewers, and borderline articles provide additional information to an SLR's readers.

6.2. Four aspects

Before comparing the outcome from the search procedure between the original SLR and the replication in more detail (or any two SLRs on the same topic), four aspects need to be separated:

1. Inclusion and exclusion criteria: Although these are stated in the original SLR, they may be interpreted differently by different research teams. Implicit agreements between the researchers involved may not be fully captured in the criteria documented, which is a challenge when comparing two SLRs or replicating an SLR.
2. Search strategy: The search may have been done in different ways, as here, when the search strategy was deliberately different, i.e., to compare the search strategies.
3. Inclusiveness: This refers to what is needed to move an article to the next assessment level. We distinguish at least three levels here. First, articles may be assessed only on their title, i.e., the immediate information available when searching in, for example, Google Scholar, or when looking at articles in a reference list. The second level concerns assessing, for example, primarily the abstract, but in conjunction with

the article's title and keywords. The third level means that the articles' full text is assessed.

4. Judgement: The final judgement concerning inclusion or exclusion is based on full–text assessment. Different researchers may not judge an article similarly since assessing an article is subjective.

The execution of the approach in the four items above-mentioned differed between the original SLR and the replication presented here. In some cases, the differences were intentional, and in other instances, unintentional.

Concerning the inclusion/exclusion criteria, we learnt, after having executed the replication, that in the original SLR, only articles having their primary focus on industry–academia collaboration in SE were included. Thus, articles describing a collaborative research effort in different areas of SE, such as requirements engineering, design or testing, and including, for example, a section on reflections concerning the collaboration were, according to one of the original SLR's authors, excluded in the original SLR. The main inclusion criterion in the original SLR was formulated as follows: "Does a given study present findings relevant for IAC in SE?" we included articles not solely focusing on industry–academia collaboration in SE. We have separated the two types of articles in the analysis to allow for a fairer comparison between the two SLRs (original and replication). We report both types of articles for two reasons. First, we perceive that the original SLR does include some articles not solely focusing on industry–academia collaboration. Secondly, we judge that some articles that do not exclusively concentrate on collaboration contain essential lessons learnt.

Second, the search strategy was intentionally different. Thus, this is the aspect we want to compare, although it is affected by the other three items above-mentioned.

Concerning the third item, after exchanging emails with the the original SLR's first author, we have taken a more inclusive approach in the replication to not miss any article due to its title and abstract not being sufficiently informative. To a large extent, our inclusiveness is a consequence of not only including articles with their primary focus on industry–academia collaboration. Thus, we have included articles with general experiences concerning industry–academia collaboration, although it is not the article's primary focus. Focusing solely on articles with the primary focus on industry–academia collaboration makes the process more efficient but increases the risk of missing essential experiences. This illustrates the influence of how the scope of

an SLR is formulated. Although not sufficiently clearly expressed in the original SLR, the more focused scoping of the original SLR explains some of the differences in the number of articles assessed by the different research teams. Whether higher inclusiveness is worth the extra effort is a subject for further research.

Finally, the judgement is likely to be different due to being done by humans and hence being subjective. Our two new concepts, i.e., wild cards and borderline articles, are intended to help mitigate the issue and make possible differences transparent. Moreover, to help understand the potential differences between the two teams (original SLR and replication), we have been provided access to articles assessed but being excluded in the original SLR.

6.3. Comparison of the original SLR and its replication

Table 4 presents a comparison of the publications listed in the original SLR and the replication. The first row (below the row with the headings) in the table highlights the number of publications included by each research team. The percentages show that the replication found 30% more relevant publications than the original SLR. It is worth mentioning that the number of publications included by both the original SLR and its replication is 20 articles.

Table 4: Comparison original and replicated SLR.

Comparison	**Original SLR**	**Replicated SLR**	**Percentage**
Listed in respective SLR	33	43	30%
Remove non-peer-reviewed publications	29	43	48%
Remove articles excluded in replication	24	43	79%
Remove after assessment articles included in original and not in replication	22	43	95%
Remove articles not having a primary focus on IAC	19	36	89%

Four publications listed in the original SLR have been removed in the second row since they are not considered peer–reviewed. It includes two keynotes, one book and a contribution to a column in IEEE Software.

Concerning this comparison, the original SLR includes nine publications not included in the replication. However, five of these have been assessed when performing the replication. Two of them were included among the borderline articles, i.e., Baldassarre et al. (2013), and Morris et al. (1998).

The third article, by Connor et al. (2009), focuses on industry–academia collaboration in education. Hence it should be excluded based on information from one of the original SLR's authors. The fourth article by Rombach et al. (2008) was excluded at the abstract level. It received a score of zero from the reviewers assessing the articles in the replication. Finally, the fifth article by Lamprecht and van Rooyen (2012) was excluded after full–text reading with a score of two. The latter article addresses IPR concerns when industry and academia collaborate, and it is in the context of regulations in South Africa. Thus, the article does not address the actual collaboration. Moreover, it focuses on one aspect related to the relationship between industry and academia. It is in the context of regulations in one country; hence it was excluded. The removal of these five articles from the original SLR led to the outcome presented in the third row in Table 4.

The remaining four articles included in the original SLR and not found in the replication were assessed by the replication's research team. It resulted in the following, an article by Franch et al. (2012) was excluded with a score of one in the full–text reading, and an article by Osterweil et al. (2008) was excluded with a score of zero on the abstract level given that it discusses the impact of research on practice and not industry–academia collaboration. Thus, given that these two articles were not relevant, we obtained the results on the fourth row in the table. The other two articles were assessed as being appropriate for inclusion. They are the articles by Raschke et al. (2014) (with a score of four) and Krishnan et al. (2009) (with a score of five).

Finally, given that we were informed by one of the original SLR's authors that only articles having the main focus on industry–academia collaboration in SE were included, we looked at all articles included by both the original SLR and the replication. It resulted in removing three articles in the original SLR and seven articles in the replication. The results from this removal can be found in the fifth row of Table 4. However, we believe that some of the articles not primarily focused on industry–academia collaboration but including essential experiences (typically in a separate section of the article) from collaboration, should be included. We find the comparison on the fourth row most relevant based on this. It includes 22 peer-reviewed articles in the original SLR and 43 peer-reviewed articles in the replication. Given the overlap of 20 articles, the "superset" contains 45 peer-reviewed articles, including essential findings concerning industry–academia collaboration published in 2014 or earlier. In summary, the targeted findings, as described in the original SLR, include collaboration models, challenges, and

best practices.

6.4. Comparison of the hybrid search strategies

When comparing the four alternative hybrid search strategies (see Section 3.1), it becomes evident that their performance is highly dependent on the start set, see Table 5. We found 43 articles when using the first alternative strategy, as presented above. The other three alternatives aim at making the work more efficient by not performing all forward and backward snowballing combinations.

Table 5: Comparison of hybrid search strategies.

Search strategy	**Number of articles**
Strategy 1: Scopus + BS*FS	43
Strategy 2: Scopus + BS||FS	23
Strategy 3: Scopus + BS+FS	38
Strategy 4: Scopus + FS+BS	23

Based on the outcome presented in Table 5, we made the following observation. In this case, most articles are found in the first round of backward snowballing and the second round of forward snowballing. Thus, running backward and forward snowballing in parallel means not benefiting from each other. When having a round with a few articles in backward or forward snowballing, we risk that the procedure stops early. Given that many articles, in our case, were found with backward snowballing in the first round, the best option is to use search strategy 3 since the articles found in backward snowballing would be used when performing forward snowballing. Thus, in our case, search strategy 3 is superior if trying to make the search more efficient.

However, as indicated above, the different hybrid search strategies' performance depends on the start set and publication patterns over the years in the investigated time interval. In our case, we created a start set by searching in Scopus in the time interval 2010-2014. Overall, we are interested in articles published in 2014 and earlier. Therefore, the search in Scopus is focused on relatively new articles published in 2014 or earlier. Thus, it is no surprise that backward snowballing is superior in the first round. Then, as we identify older articles, forward snowballing performs well since it looks at newer articles relative to those found in backward snowballing. It indicates that alternating between backward and forward snowballing may be an option (in this case). If starting with one round of backward snowballing

and then continuing with one round of forward snowballing and continuing alternating between the two ways of performing snowballing, we identify 40 articles. It does not mean that we find all 43 articles, but more articles than the hybrid search strategies 2–4.

A potential way forward is to have an adaptive hybrid search strategy depending on the start set. In this case, the adaptation would be to have an alternating hybrid search strategy since we started with relatively few articles in the start set, and hence going backwards first is most likely the best option. We may also consider the number of references in relation to the number of citations to the articles in the start set to decide how to make the searches more efficient. However, this is an area for future research.

6.5. Reflections on the evaluation

We chose to evaluate the hybrid search strategy by performing a replication to have a point of reference. The objective was to compare the outcome from the hybrid search strategy with a search performed using another search strategy. In addition to the validity threats discussed in Section 7, some aspects are essential to reflect upon:

- Different researchers – The same researchers cannot perform a replication using a different search strategy unless there is a substantial time gap between the two searches. Thus, replication is typically done by new researchers. However, it comes with challenges. Different researchers may have different interpretations of a topic as discussed by Wohlin et al. (2013). Furthermore, the inclusion/exclusion process is subjective, and hence it is highly likely that differences in included and excluded articles appear (cf. Section 6.2).

- Quality of SLR – When it comes to the quality, it is not primarily the quality of the SLR as such, but the quality of the search that is essential here. The selected SLR is published in the Information and Software Technology journal, which ought to imply a certain level of quality. Furthermore, our impression when reading the SLR, particularly the parts related to the search strategy and the inclusion/exclusion process, was that it was a well–conducted study. As the replication progressed, several questions arose concerning the process, as some information was missing. However, the authors of the original SLR were very supportive

with additional information after our search was completed. In summary, it may be concluded that even if a study seems well–described, a replication tends to reveal missing details.

- Quality of outcome – The quality of the outcome, in terms of articles identified, depends on both the researchers and the process' documentation. Thus, having another set of researchers performing a replication and using a different search strategy will likely generate a slightly different set of articles. However, different replications may still come to the same conclusions concerning the evidence in the SLR's area.

- Effectiveness and efficiency – According to the guidelines by Kitchenham and Charters (2007), we should include all primary studies related to the research questions. However, it is never the case since we are always bound to restrict ourselves, for example, concerning the language. The guidelines indicate that effectiveness is more important than efficiency by stating that we should identify all primary studies. In practice, it is a different matter. On the one hand, the hybrid search strategy includes looking at both reference lists and citations, which may result in a substantial number of articles to assess. On the other hand, Garousi et al. (2016) used 54 search strings and searched in three databases or indexing services. Given that neither of the two teams kept track of the time for performing the assessment, it is impossible to know which search strategy was most efficient. However, the hybrid search strategy was more effective given that more articles were identified.

- Publication bias – A potential concern for all SLSs is that some primary studies' authors are over–represented, which risks biasing the conclusions from the study if not taking special care. It may become even more sensitive if the SLRs' authors are over–represented among the authors of primary studies. Here, the potential bias relates to the inclusion and exclusion of articles since the focus is on the search strategy and not the synthesised evidence. To further investigate the inclusion/exclusion, we looked at articles published by the researchers from both teams in the area. Seven publications authored by researchers performing the searches are identified. One publication included by Garousi et al. (2016) is a keynote by Wohlin, which is not included in the replication due to it not being peer–reviewed. The other six articles

are two articles authored by Petersen (Petersen et al. (2014), and Petersen and Engström (2014)) and four articles by Wohlin (Gorschek et al. (2006), Wohlin et al. (2012), Wohlin (2013), and Wohlin and Regnell (1999)). Garousi et al. (2016) included the journal version of Wohlin and Regnell (1999). Both the original SLR and the replication included all six articles, although in one case, two different versions. Thus, no inclusion or exclusion bias was observed.

7. Threats to validity

In this case, the threats to conclusion validity are primarily related to selection and assessor biases. The selection of a specific SLR may bias towards the hybrid search strategy. However, the SLR was selected using a set of criteria directed towards the SLR's content and did not favour a hybrid search strategy. The topic of the selected SLR was not crucial since the objective was to evaluate the search strategy and not the synthesis of evidence for the identified articles. However, it was essential that the researchers performing the replication were comfortable with the topic to ensure that articles were included respectively excluded as correctly as possible. Furthermore, the selected SLR's authors have been helpful with information concerning the original SLR, which have been valuable to ensure that our interpretation of the SLR is as good as possible.

Furthermore, there is a risk that the individual researchers become biased, given that there is a vested interest in the hybrid search strategy. However, having three researchers perform independent assessments on all articles using the new concept of wild cards helped mitigate individual assessor bias. Moreover, the new concept of borderline articles makes delimitation between inclusion and exclusion more transparent to readers. Such transparency is essential to allow readers to assess the potential assessor bias. Overall, it is judged that the study's design and the predefined criteria for selecting an SLR to use in the evaluation help minimise the conclusion validity threats.

Another potential threat in literature reviews is publication bias, i.e., articles with specific characteristics are more often published or more often retrieved. We did not assess whether publication bias favours a particular search strategy.

8. Conclusions

The overall objective was to compare and evaluate a hybrid search strategy with a search using databases and indexing services. Furthermore, we wanted to compare four alternative approaches to performing hybrid searches. To do so, an SLR was selected for replication of the search procedure. The search strategy in the original SLR was judged to be representative of how searches often are performed when doing an SLR, and it met a set of predefined criteria for selecting an SLR for replication.

When it comes to RQ1, we conclude that the full-fledged hybrid search strategy was superior to the search strategy employed in the original SLR. Only research articles published in journals, at conferences and workshops, and as book chapters were included in the replication. However, even when accepting keynotes, column contributions and books in the original SLR, the hybrid search strategy found 30% more articles than the original SLR. And it performed even better if, for example, only accepting peer-reviewed articles and removing articles assessed and excluded in the replication. In summary, the hybrid search strategy is a competitive contender as a search strategy when performing systematic literature studies.

Concerning RQ2, we conclude that the full-fledged hybrid search strategy is better than the alternative hybrid strategies. However, it requires more effort to assess all the identified articles. When comparing the full-fledged hybrid search strategy, it became clear that the success of the alternative hybrid search strategies depends on the start set, particularly relating to how it was identified. In our case, the start set includes only relatively new articles in the investigated time interval (articles published before 2015). Hence, it is no surprise that in the first round, backward snowballing found more articles for inclusion than forward snowballing, and then it is beneficial to run forward snowballing after having done backward snowballing. Thus, the third hybrid search strategy is the second best. If aiming at saving some effort, although missing some articles, it is probably best to choose an alternative hybrid search strategy based on the characteristics of the start set. Further research into adaptive (in relation to the start set) hybrid search strategies is needed.

In addition to the two research questions, two new concepts are proposed to embrace the differences in judgement when assessing articles against the inclusion and exclusion criteria. We introduced the concept of wild cards to allow for individual reviewers to put forward an article for full–text assess-

ment even if the other reviewers think the article should be excluded when assessing the title, abstract and keywords. One wild card made it into being included in the final set of articles. The second concept is borderline articles. We suggest that articles close to being included are kept in a separate list to allow for readers to judge these articles. Notably, four out of nine borderline articles resulted from being nominated as wild cards.

In summary, the full-fledged hybrid search strategy identified substantially more articles presenting models, challenges and best practices concerning industry-academia collaboration in research than the original SLR. The results strengthen the findings in Mourão et al. (2017), and Mourão et al. (2020), where it was indicated that a hybrid search strategy might be a suitable alternative to identify primary studies. Here, we conclude that the hybrid search strategy is likely a competitive alternative to other search strategies. Furthermore, we suggest that the searches for SLSs are complemented with two concepts, i.e., wild cards and borderline articles.

Further research concerning the search strategies for systematic literature studies, and, in particular, comparisons between different search strategies are needed, including the hybrid search strategy.

Acknowledgement

First, we would like to express our gratitude to Prof. Vahid Garousi and Prof. Kai Petersen for providing additional details concerning their systematic literature review on industry-academia collaboration.

Furthermore, we are grateful to Dr. Frank Houdek, Dr. Teade Punter and Dr. Piërre van de Laar for helping us get access to articles we were unable to obtain without the authors' help.

Professor Marcos Kalinowski is funded by a research grant from the Brazilian National Council for Scientific and Technological Development (CNPq), Grant #312827/2020-2.

Appendix A. Supplement – Selecting an SLR for replication

Appendix A describes how we selected an SLR to replicate.

Appendix B. Supplement – Included articles and borderline articles

Appendix B provides listings of the included articles and the borderline articles, respectively. The listings show in which step of the hybrid search each article was identified.

References

Agnew, N.M., Pyke, S.W., 2007. The Science Game – An Introduction to Research in the Social Sciences. 7th ed., Oxford University Press.

Badampudi, D., Wohlin, C., Petersen, K., 2015. Experiences from using snowballing and database searches in systematic literature studies, in: Proceedings International Conference on Evaluation and Assessment in Software Engineering, p. 17.

Baldassarre, M.T., Caivano, D., Visaggio, G., 2013. Empirical studies for innovation dissemination: Ten years of experience, in: Proceedings International Conference on Evaluation and Assessment in Software Engineering, pp. 144–152.

Connor, A.M., Buchan, J., Petrova, K., 2009. Bridging the research-practice gap in requirements engineering through effective teaching and peer learning, in: Proceedings International Conference on Information Technology: New Generations, pp. 678–683.

Dieste, O., Grimán, A., Juristo, N., 2009. Developing search strategies for detecting relevant experiments. Empirical Software Engineering 14, 513–539.

Felizardo, K.R., Mendes, E., Kalinowski, M., Souza, E.F., Vijaykumar, N.L., 2016. Using forward snowballing to update systematic reviews in software engineering, in: Proceedings International Symposium on Empirical Software Engineering and Measurement, pp. 1–6.

Felizardo, K.R., da Silva, A.Y.I., de Souza, E.F., Vijaykumar, N.L., Nakagawa, E.Y., 2018. Evaluating strategies for forward snowballing application to support secondary studies updates: Emergent results, in: Proceedings Brazilian Symposium on Software Engineering, pp. 184–189.

Franch, X., Ameller, D., Ayala, C.P., Cabot, J., 2012. Bridging the gap among academics and practitioners in non-functional requirements management: Some reflections and proposals for the future, in: Seyff, N., Koziolek, A. (Eds.), Modelling and Quality in Requirements Engineering: Essays Dedicated to Martin Glinz on the Occasion of His 60th Birthday. Verlagshaus Monsenstein und Vannerdat, Muenster, pp. 267–273.

Garousi, V., Petersen, K., Ozkan, B., 2016. Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review. Journal of Information and Software Technology 79, 106–127.

Gorschek, T., Garre, P., Larsson, S., Wohlin, C., 2006. A model for technology transfer in practice. IEEE Software 23, 88–95.

Jalali, S., Wohlin, C., 2012. Systematic literature studies: Database searches vs. backward snowballing, in: Proceedings 6th International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 29–38.

Kitchenham, B.A., Brereton, P., Turner, M., Niazi, M.K., Linkman, S., Pretorius, R., Budgen, D., 2010. Refining the systematic literature review process–two participant-observer case studies. Empirical Software Engineering 15, 618–653.

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.

Krishnan, P., Ross, K., Pari-Salas, P., 2009. Industry academia collaboration: An experience report at a small university, in: Proceedings Conference on Software Engineering Education and Training, pp. 117–121.

Lamprecht, S.J., van Rooyen, G.J., 2012. Models for technology research collaboration between industry and academia in South Africa, in: Proceedings IEEE Software Engineering Colloquium (SE), pp. 11–17.

MacDonell, S., Shepperd, M., Kitchenham, B., Mendes, E., 2010. How reliable are systematic reviews in empirical software engineering? IEEE Transactions on Software Engineering 36, 676–687.

Mendes, E., Felizardo, K., Wohlin, C., Kalinowski, M., 2019. Search strategy to update systematic literature reviews in software engineering, in: Proceedings Euromicro Conference on Software Engineering and Advanced Applications, pp. 355–362.

Morris, P., Masera, M., Wilikens, M., 1998. Requirements engineering and industrial uptake, in: Proceedings International Symposium on Requirements Engineering, pp. 130–137.

Mourão, E., Kalinowski, M., Murta, L., Mendes, E., Wohlin, C., 2017. Investigating the use of a hybrid search strategy for systematic reviews, in: Proceedings International Symposium on Empirical Software Engineering and Measurement, pp. 193–198.

Mourão, E., Pimentel, J.F., Murta, L., Kalinowski, M., Mendes, E., Wohlin, C., 2020. On the performance of hybrid search strategies for systematic literature reviews in software engineering. Journal of Information and Software Technology 123, 106294.

Osterweil, L.J., Ghezzi, C., Kramer, J., Wolf, A.L., 2008. Determining the impact of software engineering research on practice. Computer 41, 39–49.

Petersen, K., Engström, E., 2014. Finding relevant research solutions for practical problems: The SERP taxonomy architecture, in: Proceedings International Workshop on Long-Term Industrial Collaboration on Software Engineering, Association for Computing Machinery, New York, NY, USA. pp. 13–20.

Petersen, K., Gencel, C., Asghari, N., Baca, D., Betz, S., 2014. Action research as a model for industry-academia collaboration in the software engineering context, in: Proceedings International Workshop on Long-Term Industrial Collaboration on Software Engineering, pp. 55–62.

Porte, G., 2013. Who needs replication? CALICO Journal 30, 10–15.

Punter, T., van de Laar, P., 2010. Industrial impact and lessons learned, in: Van de Laar, P., Punter, T. (Eds.), Views on Evolvability of Embedded Systems. Springer, Dordrecht, pp. 279–299.

Raschke, W., Zilli, M., Loinig, J., Weiss, R., Steger, C., Kreiner, C., 2014. Embedding research in the industrial field: A case of a transition to a

software product line, in: Proceedings International Workshop on Long–Term Industrial Collaboration on Software Engineering, pp. 3–8.

Rombach, D., Ciolkowski, M., Jeffery, R., Laitenberger, O., McGarry, F., Shull, F., 2008. Impact of research on practice in the field of inspections, reviews and walkthroughs: Learning from successful industrial uses. SIGSOFT Software Engineering Notes 33, 26–35.

Shull, F.J., Carver, J.C., Vegas, S., Juristo, N., 2008. The role of replications in empirical software engineering. Empirical Software Engineering 13, 211–218.

Skoglund, M., Runeson, P., 2009. Reference-based search strategies in systematic reviews, in: Proceedings International Conference on Evaluation and Assessment in Software Engineering, pp. 31–40.

Wohlin, C., 2013. Empirical software engineering research with industry: Top 10 challenges, in: Proceedings International Workshop on Conducting Empirical Studies in Industry, pp. 43–46.

Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings International Conference on Evaluation and Assessment in Software Engineering, pp. 321–330.

Wohlin, C., 2016. Second-generation systematic literature studies using snowballing, in: Proceedings International Conference on Evaluation and Assessment in Software Engineering, pp. 1–6.

Wohlin, C., Aurum, A., Angelis, L., Phillips, L., Dittrich, Y., Gorschek, T., Grahn, H., Henningsson, K., Kagstrom, S., Low, G., Rovegard, P., Tomaszewski, P., van Toorn, C., Winter, J., 2012. The success factors powering industry-academia collaboration. IEEE Software 29, 67–73.

Wohlin, C., Mendes, E., Felizardo, K.R., Kalinowski, M., 2020. Guidelines for the search strategy to update systematic literature reviews in software engineering. Information and Software Technology 127, 106366.

Wohlin, C., Regnell, B., 1999. Achieving industrial relevance in software engineering education, in: Proceedings Conference on Software Engineering Education and Training, pp. 16–25.

Wohlin, C., Runeson, P., da Mota Silveira Neto, P.A., Engström, E., do Carmo Machado, I., de Almeida, E.S., 2013. On the reliability of mapping studies in software engineering. Journal of Systems and Software 86, 2594–2610.

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