IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 6 June 2023, accepted 18 July 2023, date of publication 26 July 2023, date of current version 14 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299211

RESEARCH ARTICLE

Reproducible Searches in Systematic Reviews: An Evaluation and Guidelines

ZHENG LI, (Senior Member, IEEE), AND AUSTEN RAINER
School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT9 5BN Belfast, U.K.

Corresponding author: Zheng Li (zheng.li@qub.ac.uk)

ABSTRACT ***[Context:]*** The Systematic Review is promoted as a more reliable way of producing a high-quality review of prior research. But there are a range of threats that can undermine the reliability and quality of such reviews. One threat is the reproducibility of automated searches. ***[Objectives:]*** To evaluate the state-of-practice of reproducible searches in secondary studies, and to consider ways to improve the reproducibility of searches. ***[Method:]*** We re-run the searches of 621 secondary studies and analyse the outcomes of those (attempted) re-runs. We use the outcomes, and our experience of re-running the searches, to propose ways to improve the reproducibility of automated searches. ***[Results:]*** With the 621 studies, more than 50% of the literal search strings (ignoring other settings) are not reusable; about 87% of the searches (e.g., with settings) cannot be repeated; and around 94% of the searches (including all elements of the search) are irreproducible. We propose guidelines for automated search, directing particular attention at the formulation of search strings. ***[Conclusion:]*** While some aspects of automated search are beyond the direct control of researchers (e.g., variations in features, constraints and performance of search engines), many aspects can be effectively managed through more careful formulation and execution of the search strings themselves, and of the search settings. While the results of our evaluation are disappointing there are many simple, concrete steps that researchers can make to improve the reproducibility of their searches.

INDEX TERMS Automated search, evidence based software engineering, reproducibility, search engine, secondary study, systematic review.

I. INTRODUCTION

Systematic reviews – notably, the systematic literature review (SLR [1]) and the systematic mapping study (SMS [2]) – are promoted as more reliable ways to achieve higher-quality reviews (e.g., [3]) of prior research. Unfortunately, there are a wide range of threats that can undermine the reliability and subsequent quality of such reviews [4], [5], [T12] (citations prefixed with T are tertiary reviews; see Section IV). Thus, the community needs to explore strategies to mitigate these threats.

A particular aspect of systematic reviews, *searching* for candidate primary studies, has been identified as one of the most problematic parts of the whole review process [6]. Searching is the most crucial stage in the evidence dataflow [7] because the results of the search provide the foundation for the subsequent review. On that basis, we think it essential to better understand threats to the reliability of searching, and to attempt to address those threats.

When searching for literature, there are two broad search approaches: *automated search* and *manual search*. With automated search, the search is still typically initiated by a manual operation (e.g., the researcher interacts with the interface of an online search engine) but the search itself is undertaken automatically by the respective *search source*. (We use the term "search source" to collectively refer to searchable content providers, such as the ACM Digital Library, and searchable content indexers, such as Google Scholar). By contrast, a manual search comprises (almost) entirely manual operations, e.g., browsing the reference section of an article. In principle, automated search brings huge economies of scale for the researcher, as well as coverage; they can quickly identify a smaller, more relevant subset of articles from a larger, more comprehensive set of candidate articles.

The associate editor coordinating the review of this manuscript and approving it for publication was Claudia Raibulet.

With the continual growth in software engineering (SE) publications, and the need to remain informed of advances in research, the researcher is increasingly dependent on automated searches.

It is within this context that we turn to the problem of searching for candidate primary studies as part of a systematic review. Automated searches are a vital part of systematic reviews. But there remain problems with the reliability of automated searches. For this article, we focus on a particular aspect of reliability, i.e., *reproducibility* of searches. We set two objectives:

For objective #1, we seek to understand the degree to which a researcher could, at some future point in time, sufficiently reproduce the results of a previously conducted search. To scope our objective, we focus on secondary studies that are clearly based on the SLR protocol [1]. We investigate the following research question: to what extent can we reproduce automated searches from existing SE secondary studies? Our investigation constitutes an evaluation of the state of practice of secondary studies in SE.

For objective #2, we seek ways to improve the reproducibility of searches. Again, we ask a question, though formally it is not a *research* question: how can we improve the reproducibility of automated search in SE secondary studies? (We do not treat this second objective as a research question because we are not investigating the world as-is, but instead exploring ways to change the world.) To achieve this objective, we formulate a set of guidelines derived from our evaluation and from our experience of conducting that evaluation.

Our article extends a preliminary evaluation [8] in several directions: we double the sample size of secondary studies, increase the number of search sources, perform a deeper analysis of the results, and propose a set of evidence-based guidelines.

Overall, our article makes the following contributions:

1) We perform a large-scale empirical evaluation of the state-of-practice for a fundamental aspect of secondary studies in SE, i.e., the formulation and execution of search strings. Whilst there have been prior evaluations, our empirical evaluation is the first study (to the best of our knowledge) to *replicate* the prior searches of a large sample of previous studies.
2) We concentrate on operational and technical aspects of searches, e.g., the impact of specific search-string formulations on search source results.
3) We identify researcher practices in search-string formulation that then "cause" (ir)reproducibility of search strings. Again, no prior research (to the best of our knowledge) has investigated these practices and their influence on the reproducibility of automated search.
4) We formulate a set of guidelines for addressing these "mis-practices". Previously-created guidelines and checklists are typically intended for the evaluation of already-published studies. Our checklist is intended to be used in the formulation, execution, and documentation of (future) searches.

The remainder of our article is organised as follows: Section II briefly reviews prior work; Section III presents the conceptual framework we use for our evaluation; Section IV describes the design of our evaluation; Section V reports the results of our evaluation; Section VI presents our guidelines; Section VII considers threats to the validity of our study; finally, Section VIII briefly reviews our objectives and contributions, and proposes directions for future research.

II. BACKGROUND AND RELATED WORK

In this section, we first review background research on the challenges of searching search sources. We then focus our review on related work, distinguishing our study from that work.

Prior literature generally frames the problems of automated search in terms of incomplete reporting (e.g., missing details) and technical limitations. For example, Kruger et al. [9] focus on the reporting of search strings, while Budgen et al. [T2] emphasise the time range, and the date, of the search. Neither of these examples highlight the other reproducibility-critical information, e.g., zonal settings [10] like title, abstract, and keywords. Although these researchers have suggested remedies – e.g., "fully document the search process" [T2] and "report more detailed information" [9] – the suggestions remain coarse-grained and can still allow for ambiguity, and thus uncertainty, about the information that should be reported and the level of detail of such reports.

Furthermore, reported search strings may not be practically reusable by others. Some "reporting issues" are actually flaws in the *formulation* of the search strings (e.g., syntax errors or format mistakes) rather than problems in the reporting of those strings. For example, with one secondary study (and similar behaviours occurred for other studies), we found it was obvious that IEEE Xplore had returned an unexpectedly high number of hits, i.e., almost three million hits from IEEE Xplore, compared to several hundred hits from the other search sources. But the authors of the original study still reported the result, and even then designed a special selection strategy, without examining their original search string (that had syntax errors) or comparing it against IEEE Xplore's search features. With such cases, although all the search details are reported, and even if we can reproduce the (flawed) searches to obtain the same results, there is no clear benefit in doing so.

In terms of technical limitations, an early tertiary study [11] found that existing search sources are not suited to supporting systematic reviews in the SE domain. Although there have been promising improvements since then, different search sources still vary significantly in size, scope, underlying model, user interface, syntax and filtering mechanisms [5], [12]. Such differences and inconsistencies threaten effective automated search processes from study to study.

In fact, some constraints (e.g., different maximum limits of search terms, operators, or characters) can require well-designed search strings to be modified; also, it is not a trivial task to mix major terms and synonyms properly [6]. This in turn results in another frequent reporting issue, i.e., the adapted search strings are missing. Since technical problems with search engines are beyond the control of researchers, a natural suggestion is to report problems to the library owners to (hopefully) fix them [9]. However, fixing all the engine constraints is an ideal and researchers will still need to confront constraints. By performing feature analyses of major digital libraries, one study [10] developed advice for researchers on how to manage existing constraints and inconsistencies in search sources. Similarly, we also show, with our guidelines later in this paper, that researchers can bypass at least some technical limitations by careful formulation and execution of their searches.

Turning to related work, the closest research to ours is a checklist, by Ali and Usman [13], to evaluate the reliability of the *reported* searches. Although there are overlapping concerns between Ali and Usman's [13] study and our study – e.g., the search string's engine-specific adaptations should be documented – we claim two main differences for our study:

- The research *method* is different: Ali and Usman [13] *aggregate* existing guidelines on searches to formulate their checklist, whilst we *replicate* 621 prior searches to develop our guidelines.
- The research *purpose* is different: Ali and Usman [13] intend for their checklist to be used to evaluate published studies, whilst we intend for our guidelines to be used in the formulation, execution and documentation of (future) searches.

Moreover, a unique feature of our study is its concentration on operational and technical aspects of searches. Unlike existing guidelines and checklists, we are not concerned with the research context of the searches. For example, we do not consider the "fitness-for-purpose" of the search string in relation to the corresponding research question. This is out of the scope of our study and, in any case, has already been emphasised extensively by existing guidelines [1], [2], [14], [15]. In contrast, the operational and technical details "are rarely communicated or documented" [9]. Thus, researchers could continuously encounter those problems even without being aware of them. Overall, to the best of our knowledge, our study is the first study that directly addresses these operational and technical aspects.

III. MODELLING THE SEARCH PROCESS

In this section, we first introduce a general model of the search process and then discuss three indicators of search, and their components, as well as our proxy measure for reproducibility. The model, indicators, components, and proxy measure provide the conceptual foundation for our evaluation.

A. A GENERAL MODEL FOR CONDUCTING SEARCHES

Figure 1 presents a model of the search process. This model is intended to be independent of platform-specific implementations, e.g., agnostic to different interfaces. Also, although we ideally want the entire search process to be automated, there are still manual operations required for searching sources. In brief, the figure indicates that a search string is formulated and adapted to a specific interface. Then, the time and the zonal settings, if any, are required, in order to restrict the search, e.g., to particular parts of articles, such as title, abstract, keywords, full text [10]. Depending on the engine's features, pre-filtering and/or post-filtering offer other ways to constrain the search (e.g., publication type, or research domain) to narrow down the search results. Finally, after executing the search, search results will be obtained. These results might then be further analysed, e.g., through the manual application of exclusion criteria, such as duplicate results.

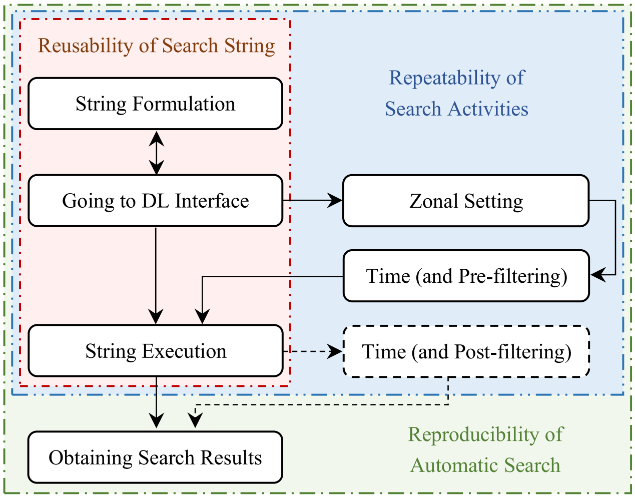

Reusability of Search String
Repeatability of Search Activities
String Formulation
Going to DL Interface
Zonal Setting
Time (and Pre-filtering)
String Execution
Time (and Post-filtering)
Obtaining Search Results
Reproducibility of Automatic Search

FIGURE 1. The generic automated search workflow for systematic reviews. (*A reusable search string = An available & executable search string + The same search sources*), (*Repeatable search activities = A reusable search string + The same zonal settings & time range*), and (*A reproducible automatic search = Repeatable search activities + The same search results*).

B. INDICATORS AND THEIR COMPONENTS

Reproducible research allows another researcher (or the same researcher at a different time) to use the available data and code to obtain the same results [16]. Applying this principle to searches, we identify several components of searches which can be treated as data and code, and which are necessary for reproducible research. These components are: available search string (e.g., is the search string stated?), executable search string (e.g., can a search be executed with the string?), search source (e.g., are the same sources being used for the searches?), zonal settings, time settings, and results from the search (e.g., are the same search results produced?). There is an additional component, filters, which we discuss in the context of the indicators.

We map the six components to three indicators: reusable search strings, repeatable searches and reproducible searches. Our mapping is summarised in Table 1. For clarity, we also include the filter component. We discuss the three indicators,

TABLE 1. Summary of search indicators.

Indicator	Reported	Executable	Same search source	Zonal	Time	(Filter[1])	Same results
Reusable string	X	X	X				
Repeatable search	X	X	X	X	X	(X)	
Reproducible search	X	X	X	X	X	(X)	X

[1] We recognise that filtering is used in searches but do not treat it as essential for our study.

their mappings, and a proxy measure for reproducibility in the following subsections.

C. A PROXY MEASURE OF REPRODUCIBILITY

For the final component, i.e., same search results, it is impossible to verify whether retrieved papers in a subsequent secondary study are identical to those in the original secondary study. Consider, for example, that no researcher (to the best of our knowledge) has reported the content, as opposed to the counts, of their search results; and it is generally unnecessary to do so, for secondary studies. Also, as time passes, the number of articles stored in the search source will likely increase between the original search and the subsequent search. Expecting a (near) identical number of results would be unrealistic.

We therefore take a pragmatic approach and compare the original search with our search in terms of a ''tolerated'' number of results, i.e., differences within an order of magnitude are tolerated as sufficient for a reproducible search. We tolerate an order of magnitude difference for three reasons. First, several search sources can have day-to-day differences in search results from the same automated search [9]. The variation in day-to-day results from search sources is generally trivial [9]. Second, publishing or indexing delays can make search results vary, when searching at different times [T2] though such delays have a limited time lag (about three months according to [T2]). Third, as already stated, the number of articles stored in the search source will likely increase over time.

D. REUSABLE SEARCH STRINGS

We treat search strings as the most critical data in the automated search workflow (see Figure 1). We define a *reusable string* as a search string that has been reported in a paper, and is therefore available to be reused, and capable of being processed with default settings in any flexible console (e.g., the command search window) of the originally reported search sources. For the purpose of fairness in the comparison between a secondary study and our searches, we only consider the search sources employed in the original study.

Compared with related work [9], our definition of a *reusable string* emphasises the practical *re-use* of a search string, rather than only checking that the string was reported in a publication. Essentially, a *reusable string* is a previously used search string that can still be reused.

E. REPEATABLE SEARCHES

Search strings may be reusable but no longer operate in the way intended. For example, there may have been changes in other aspects of the search engine. The next step in verifying the reproducibility of a search is therefore to consider the zonal and time settings of the search. We define a *repeatable search* as a *reusable search string* that can be processed by the reported search source with exactly the same zonal settings and time range as with the original search.

If further filtering details (e.g., filtering Subject Area) are reported with the original search, we try to repeat that filtering with our search, to be fair in our comparison. Considering the large diversity of filters for the different search sources, we do not treat filter settings as a required criterion for this indicator. On the other hand, since unrepeatable or unknown filter settings may significantly influence the search results, such negative influences should still be capture-able by our final indicator, *reproducible search*.

F. REPRODUCIBLE SEARCHES

We define a *reproducible search* as a *repeatable search* that also produces sufficiently similar search results. Provided the new results are within one order of magnitude difference to the original results, we consider the search to be reproducible. As an example, we consider Novais et al.'s study [17] to have reproducible searches, even though our test on IEEE Xplore obtains almost twice as many hits as the reported amount (i.e., 1651 vs. 865) as the results are within one order of magnitude difference. By implication, a reproducible search has either repeated the filtering settings of the original, or the filtering settings did not have a sufficiently different effect on the searches.

IV. METHOD OF EVALUATION

In this section we explain the method we used to sample secondary studies. Essentially, we *snowballed* [19] from tertiary studies. We then discuss our methods of analysing the secondary studies.

A. SEARCHING FOR TERTIARY STUDIES

Since its introduction, by Kitchenham and Charters in 2004 [1], the SLR methodology has been widely employed in SE. Given the ongoing growth in the number of SLRs over the last twenty years, it is difficult to gather the population of SE secondary studies. We therefore apply the *snowballing* [19] approach to a selection of tertiary studies of SE secondary studies, to identify a relatively large number of secondary studies of SE. By using tertiary studies of SE secondary studies, we can be more confident in the relevance, and therefore the representativeness, of our sample.

To identify a candidate set of tertiary studies, we follow the short-string strategy [20] to maximise the search scope. We use the following search string:

"tertiary study" AND "software" AND "review"

In our previous study [8], we relied on Google Scholar as the single source to search for candidate tertiary studies published in the past five years. We chose "sort by relevance" to limit our screening to the first 20 pages of the search results (10 results per page). Our decision to screen the first 20 pages is similar to previous research, e.g., [21]. We then reviewed the titles and snippets of text returned by Google Scholar.

For the current study, we considerably extend our sources, searching the four major digital libraries for SE, as advocated by Zhang and Babar [22]:

- ACM Digital Library: https://dl.acm.org/
- IEEE Xplore: https://ieeexplore.ieee.org/
- ScienceDirect: https://www.sciencedirect.com/
- SpringerLink: https://link.springer.com/

We used the same search string to conduct all-field searches across the four digital libraries (in SpringerLink, we used the default search within discipline Computer Science) for the period 2015–2020 inclusive. Since tertiary studies need secondary studies, and given that the SLR guidelines were *first published* in 2004, restricting our search to 2015 ensures we find tertiary studies that have, in principle, reviewed up to ten years of published secondary studies. Our searches of the four digital libraries respectively returned 37, 52, 123 and 42 results (before applying exclusion criteria).

B. SELECTING RELEVANT TERTIARY STUDIES

When selecting relevant tertiary studies, we predefined a set of inclusion and exclusion criteria, as specified in the upper half of Table 2. Using two studies as examples, we briefly note here one unfortunate constraint: Batouta et al. [T13] selected 2347 secondary studies, but only report 11 in their paper. We can therefore only use the 11 ($<1\%$) of their secondary studies. And Bayram et al. [23] refer to a dataset of 94 SLRs, but none of these are specified in their (short) paper. Thus, we cannot use any of their secondary studies.

As we do not need to comprehensively verify the research details at this stage (such verification comes later), we briefly reviewed the candidate publications and identified a total of 24 relevant tertiary studies (see Table 3) in the SE domain. The 11 studies identified in our previous study [8] are all present in the 24 studies for the current study. (This further confirms the benefits of a multi-source search strategy [1], [11]. Google Scholar's broad coverage may be at the expense of weak precision.)

C. COLLECTING SAMPLES OF SECONDARY STUDIES

From the 24 tertiary studies, we identified 1326 candidate secondary studies. After removing duplicates and applying exclusion criteria, summarised in the lower half of Table 2, we finally selected 621 secondary studies. (For consistency, we excluded four papers that had been considered in our previous study [8]. These four papers had not been peer-reviewed.) A summary of the tertiary papers and the finally-selected primary studies is presented in Table 3. The complete list of included and excluded studies is available at `http://doi.org/10.5281/zenodo.4447488`. Given our method for selecting secondary studies via tertiary studies, and of then filtering the candidate set of secondary studies, our sample of 621 studies should be of higher quality.

D. ANALYSING THE PREVALENCE OF OUR INDICATORS AND SEARCH SOURCES

Having selected 621 secondary studies, we extracted the search string for each study. To support like-to-like comparison, we only analyse one string for each study. If there exist multiple search strings in a study, we select the longest one for our comparison. Many studies report only one string. Some studies report a "standard string" which has then been adapted, but (all of) the adaptions were not reported. Thus, we needed to "normalise" the studies for comparison.

Then, for each of the 621 strings, we retried the search string, collecting information on the success of the resulting search relative to our three indicators (see Section III and Table 1) and on the search sources used. We report the results in Section V-A and Section V-B.

E. STUDYING THE CAUSES OF PREVALENCE

During our tests of the search strings and searches, we frequently observed contrasting practices in terms of the search string length and the search source amount. For example, some researchers tend to use exhaustive search terms (e.g., [24]) and to use as many search sources as possible (e.g., [25]), while some others prefer to use less keywords to enlarge the search scope (e.g., [20]) and to use limited indexers (or even a single one) so as to have a broad coverage of literature (e.g., [26]). To the best of our knowledge, no research has compared these different practices and investigated their influences on automated search.

By treating each search as an independent experimental trial, we can design factorial ANOVAs [27] to study potential explanations for the (lack of) success in our searches. We identified three factors and constructed ANOVAs for our three indicators. The three factors and their levels, giving a 2^3 factorial ANOVA, are: noitemsep

- **Venue type**: Journal vs. Conference
- **String term-count**: Long (>11 search terms) vs. Short (otherwise)
- **Number of search sources**: Many (>5 search sources) vs. A few (otherwise)

TABLE 2. Inclusion and exclusion criteria for tertiary and secondary studies.

ID	Criterion and explanation
Inclusion criteria for tertiary studies	
IT1	Peer-reviewed journal or conference paper: A peer-reviewed tertiary study is more likely to be of higher quality, and to use higher-quality secondary studies.
IT2	Explicitly shares all the selected secondary studies: We need the secondary studies for our sample.
IT3	Shares some of the selected secondary studies: We can use the subset of secondary studies in our sample.
Exclusion criteria for tertiary studies	
ET1	Publication is a secondary study: To avoid confusion, secondary studies are ignored when choosing tertiary studies.
ET2	Not in the SE domain: Publications outside the SE domain are not relevant for our study.
ET3	Publication has not reported its selected secondary studies: Although it is possible to request datasets from authors, we choose not to do so, partly for pragmatic reasons (such as time), partly for transparency (such datasets are not publicly available for others to examine) and partly for consistency (e.g., some authors may decline, others may agree).
ET4	Selected secondary studies are inaccessible: A tertiary study may refer to a repository, or similar, of the secondary studies but the repository is not (or no longer) accessible.
Inclusion criteria for the secondary studies	
IS1	Automated, search-based secondary study: Clearly, these are the type of study we require for our evaluation.
IS2	Secondary study with a mixed search strategy: We define a mixed search strategy as a combination of manual search and automated search. The automated search part of the original studies is within scope for our evaluation.
IS3	Venue-specific secondary study: A venue-specific search can either be completely manual or using publication venues as automated search filters. The former case is excluded by other exclusion criterion ES5; the latter case is within scope for our evaluation.
IS4	Mixed-methods study involving SLR: Similar to IS2, the original study can partially fit in the scope of this research, provided their SLR implementations employs automated search.
Exclusion criteria for the secondary studies	
ES1	Published before the year 2004: The SLR protocol was first published in 2004. Thus, reviews before 2004 are out of scope.
ES2	Traditional surveys: Traditional surveys typically do not report a search process. Even when they do report a process, this process is unlikely to conform to the SLR protocol.
ES3	Student theses or dissertations: These are a particular sub-type of ES8.
ES4	Technical reports: Technical reports are not peer-reviewed (see ES8) so formally we exclude them, however they can be used as a supplement to a peer-reviewed publication, such as providing more detail on the searches, e.g., [18]. We therefore sometimes use such reports to support our evaluation.
ES5	Secondary studies based on manual search: The manual-search strategy is outside the scope of our evaluation.
ES6	Tertiary studies: Occasionally, a tertiary study [T18] may also include other tertiary studies (e.g., [T19]). For clarity and consistency, tertiary studies cited in other tertiary studies are excluded, even where they have used automated search.
ES7	(Partially) not written in English. As well as excluding studies that are entirely not written in English, we also exclude studies that include some non-English, in particular that have search strings composed of English and non-English terms.
ES8	Not peer-reviewed journal or conference papers: We use peer-review as a criterion for research quality.

For venue type, we distinguish between conference and journal as two types of publication venues. Workshops, short papers and chapters are directly labelled as conference papers. Although chapters can be comparable in length to journal articles, many conference proceedings may have been published as chapter-based books, e.g., those by Springer.

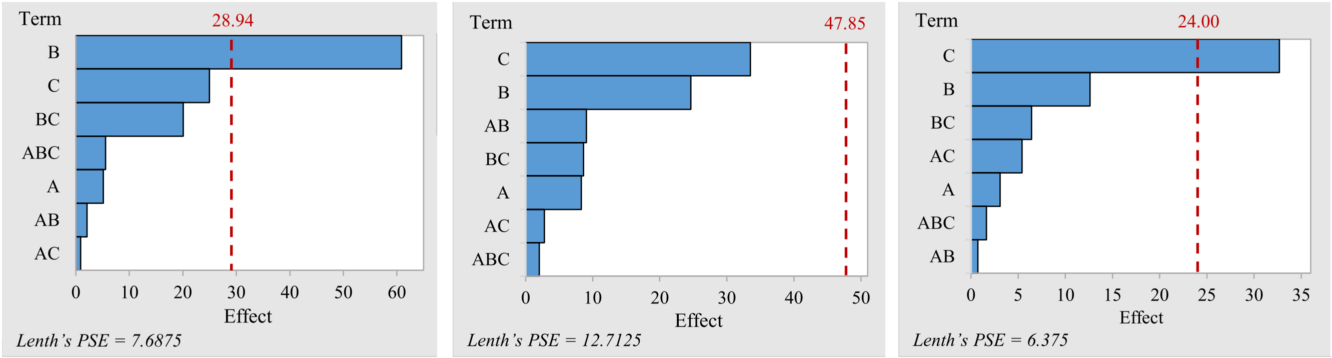
We study string term-count because of the Boolean operator limits in some search sources. We measure the length of search strings in terms of the number of terms. By counting the search terms of each string, we observe that the string term-counts of different studies vary significantly, ranging from one term to 99 terms. Since the median length is 11, we define a long string to have more than 11 search terms, otherwise the string is treated as a short string.

For the number of search sources, the usage statistics for the number of sources used (see Section V-B) suggests a median number of sources is five. Therefore, we distinguish between a secondary study employing many (i.e., more than five) search sources or a few (i.e., less than or equal to five) sources.

We calculate the rate, of the respective search indicator, as the quantitative response under each of the 2^3 factorial conditions, as specified in part (a) of Table 4. Then, we statistically investigate the effects of individual factors and of different factor interactions. Since the 2^3 factorial conditions indicate a full factorial experimental design, the effect calculation here can be formulated into Eq. (1).

$$E = \frac{\left|\sum_{i=1}^{n}(Rh_i - Rl_i)\right|}{n} \tag{1}$$

where a factor's effect (or multiple factors' interaction effect) E is represented by the average difference between responses from the factor's two alternative levels



(a) Reusable strings (b) Repeatable searches (c) Reproducible searches

FIGURE 2. Pareto charts of the effects of factors, for successful searches according to our three indicators ($\alpha = 0.05$). A=Venue; B=String term-length; C=Number of sources; AB, AC, BC, ABC = interactions of factors.

TABLE 3. Tertiary studies and counts of selected secondary studies (SS)*.

Tertiary study	# of SS	Tertiary study	# of SS	Tertiary study	# of SS
[T1]	56	[T9]	86	[T17]	9
[T2]	37	[T10]	13	[T18]	16
[T3]	14	[T11]	110	[T19]	101
[T4]	22	[T12]	165	[T20]	4
[T5]	28	[T13]	11	[T21]	70
[T6]	24	[T14]	48	[T22]	210
[T7]	60	[T15]	19	[T23]	40
[T8]	41	[T16]	4	[T24]	138

*621 valid and unique secondary studies selected.

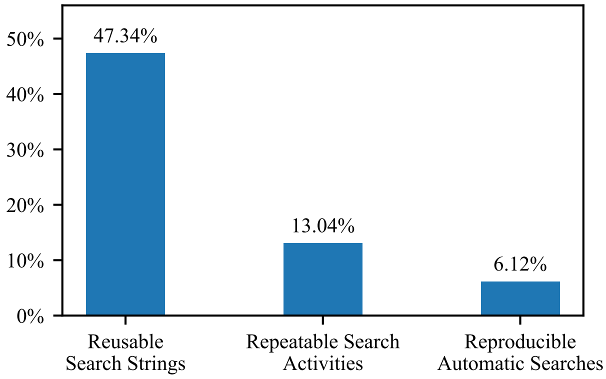


FIGURE 3. Successful searches according to our three indicators.

(i.e., high-level response *Rh* and low-level response *Rl*). For example, by referring to Table 4 and focusing on String term-length only, its effect on reusable searches is $|(0 + 51.5 + 3.4 + 42) - (88.6 + 88.8 + 76.7 + 86.3)|/4 = 60.875(\%)$.

To facilitate our analyses, we employed the DOE module of Minitab Statistical Software[1] and utilised its Pareto Chart of the Effects to illustrate the analysis result, as shown in Figure 2. Each sub-chart includes a dashed red line indicating the threshold for a statistically significant effect.

Factorial ANOVA does not directly prove any factor's effect; instead, it facilitates gaining objectivity from observations and adding objectivity to conclusions. For example, by applying such analytical results back to the original observations on string reusability rates (see Table 4), we can confidently draw a conclusion that shorter search strings have better reusability in SE secondary studies.

V. RESULTS

A. PREVALENCE OF THE INDICATORS

After retrying each search string for every secondary study, we obtained frequency counts for our three indicators. These counts are illustrated in Figure 3. The figure shows that less than 50% of our sample (294 out of the 621 studies) have *reusable* search strings. An immediate implication is that over half of the secondary studies "fail" on the most fundamental indicator for reproducible search. Given our method for selecting secondary studies via tertiary studies, our sample of 621 studies should be of higher quality. Thus, the results we report may be on the more "optimistic" side.

When applying zonal settings and time range, we can successfully conduct *repeatable searches* for about 13% of the total sample (81 studies). Finally, when comparing the search *results*, we can successfully achieve *reproducible results* for about 6% of the total sample (38 studies), even after including the studies with fixable syntax errors. Thus, approximately 95% of the secondary studies we examined are not reproducible.

B. PREVALENCE OF THE SEARCH SOURCE

There is consensus in the SE community to use multiple sources for searching, as an individual source does not contain all (or a majority) of the relevant publications. By counting how many times individual search sources are employed, we obtain their usage frequencies in the secondary study samples, as illustrated in Figure 4. Overall, 42 different sources

[1] https://www.minitab.com/en-us/products/minitab/

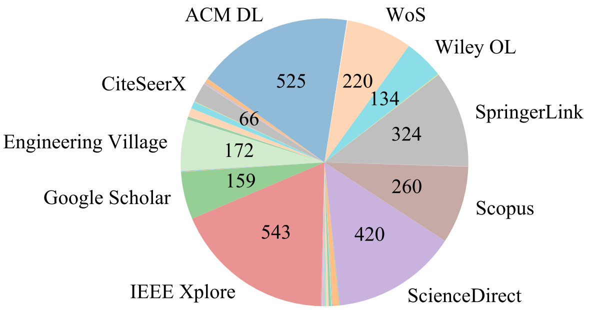


FIGURE 4. Usage frequencies of search sources in the 621 secondary studies.

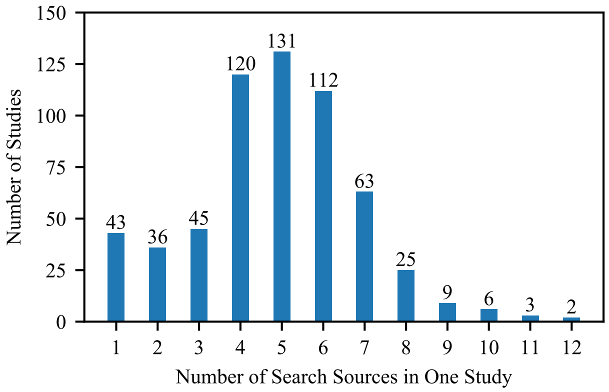


FIGURE 5. Number of search sources in secondary studies.

were used by the secondary studies;[2] for conciseness, we only label the figure with the search sources that are employed by 31 (5%) or more secondary studies. Engineering Village is the standard entrance for visiting Compendex or INSPEC, and some studies only mentioned "Engineering Village" when reporting search sources (e.g., [28]). We therefore unify the search source to be Engineering Village in both cases of Compendex and INSPEC. Similarly, we combined IEEE Computer Society Digital Library with IEEE Xplore.

By counting how many search sources are employed in one systematic review, we find that researchers tend to choose four to six search sources to explore empirical evidence, as illustrated in Figure 5. Also, by referring to Figure 4, the most popular sources are ACM Digital Librar (ACM DL), CiteSeerX, Engineering Village, Google Scholar, IEEE Xplore, ScienceDirect, Scopus, SpringerLink, Wiley Online Library (Wiley OL), and Web of Science (WoS). With the exception of CiteSeerX, which we discuss shortly, we argue that SE's major digital libraries, suggested by [22], should be updated to include this list of popular search sources. Although CiteSeerX is on the list, we do not recommend CiteSeerX, due to its strange behaviours in our tests [29]. In fact, its early version, CiteSeer, has already been identified to have inconsistent and unexplainable search behaviours [11], while the current CiteSeerX seems to have even more unpredictable behaviour.

There is ongoing debate on the use of Google Scholar in SE secondary studies. Google Scholar has a broader coverage of scientific publications than those independent publishers and indexers [30], [31]. Such a broad coverage may be particularly helpful for SE topics that involve multidisciplinary concepts across a diverse domains [32]. But the precision of Google Scholar's search results is generally low [33], [34]. This appears to be especially influenced by the "noise" from grey literature [35], which inevitably results in difficulties for researchers to manage large queries with extensive results [36]. In fact, by contrasting our use of Google Scholar in the earlier study [8] with our use of four sources in the current study, we corroborate the advice to not replace the individual search sources with Google Scholar, even though Google Scholar has "considerable overlap with ACM and IEEE on software engineering literature" [34], [37]. We do recognise the value of Google Scholar to facilitate snowballing [19] however snowballing is fundamentally a manual search, and is outside the scope of our current study. Also, as noted already, Google Scholar is beneficial for grey literature searches.

C. FACTORIAL ANOVA OF REUSABLE SEARCH STRINGS

Turning now to the factorial ANOVA, Figure 2(a) indicates that the term-count of a search string significantly effects its reusability. The figure also suggests that the number of search sources may also have an influence (also interacting with term-count), while the venue type has little effect. By applying such analytical results back to the original observations on string reusability rates (see Table 4(a)), we argue that fewer search terms increased the likelihood of *reusability* of search strings in SE secondary studies. There are recommendations to use as many as possible synonyms and alternatives of keywords (e.g., [6]). As a compromise between our recommendation and prior recommendations, researchers might perform a careful trade-off, selecting fewer, but more sensitive, search terms.

D. FACTORIAL ANOVA OF REPEATABLE SEARCHES

To assess *repeatable searches* we need information on both zonal settings and time range. Thus, we exclude those studies, and their respective search string, where information is

[2] The 42 search sources are: 1) ACM DL; 2) AIS eLibrary; 3) Australian Education Index; 4) BASE (Bielefeld Academic Search Engine); 5) Blackwell-Synergy; 6) Cambridge University Press; 7) CiteSeerX; 8) CSB (The Collection of Computer Science Bibliographies); 9) DBLP; 10) EBSCO; 11) Embase; 12) Emerald Insight; 13) Emeroteca Virtuale; 14) Engineering Village; 15) ERIC (Education Resources Information Centre); 16) Expanded Academic; 17) Google Scholar; 18) IEEE Xplore; 19) IET Digital Library; 20) IGI Global; 21) InderScience Online; 22) INFORMS PubsOnLine; 23) IOPscience; 24) JSTOR; 25) Kluwer Online; 26) Metapress; 27) Microsoft Academic Search; 28) MIS Quarterly; 29) MIT Press; 30) Oxford Journals; 31) ProQuest; 32) SAGE Journals; 33) Science (AAAS); 34) ScienceDirect; 35) Scopus; 36) SIAM (Society for Industrial and Applied Mathematics); 37) SpringerLink; 38) Taylor & Francis Online; 39) University of Hertfordshire's Library Search; 40) Wiley OL; 41) WoS; 42) World Scientific.

TABLE 4. Assessment of rates of searches under different conditions.

(a) Reusable string conditions					
Venue Type	String Length	Source Amount	Available string	Reusable string	Rate
Journal	Long	Many	72	0	0.0%
Journal	Long	A few	66	34	51.5%
Journal	Short	Many	44	39	88.6%
Journal	Short	A few	80	71	88.8%
Conf.	Long	Many	29	1	3.4%
Conf.	Long	A few	81	34	42.0%
Conf.	Short	Many	43	33	76.7%
Conf.	Short	A few	95	82	86.3%
(b) Repeatable search conditions					
Venue Type	String Length	Source Amount	Reusable strings	Repeatable search	Rate
Journal	Long	Many	36	0	0.0%
Journal	Long	A few	29	12	41.4%
Journal	Short	Many	27	12	44.4%
Journal	Short	A few	45	29	64.4%
Conf.	Long	Many	9	0	0.0%
Conf.	Long	A few	28	12	42.9%
Conf.	Short	Many	9	2	22.2%
Conf.	Short	A few	27	14	51.9%
(c) Reproducible search conditions					
Venue Type	String Length	Source Amount	Repeatable search	Reproducible search	Rate
Journal	Long	Many	23	0	0.0%
Journal	Long	A few	17	6	35.3%
Journal	Short	Many	14	3	21.4%
Journal	Short	A few	32	13	40.6%
Conf.	Long	Many	5	0	0.0%
Conf.	Long	A few	14	6	42.9%
Conf.	Short	Many	6	1	16.7%
Conf.	Short	A few	18	9	50.0%

incomplete; and use the remainder of the studies and strings to calculate the rate of repeatable searches. The results are shown in Table 4(b).

Given the reference line that indicates the statistically significant effect 47.85(%), we claim that none of the factors or factor-interactions significantly impacts the repeatability of search activities. But there are still clear differences among those factorial effects.

E. FACTORIAL ANOVA OF REPRODUCIBLE SEARCHES

For *reproducible searches*, it is meaningless to investigate any factorial effect where the original results are unavailable for comparison. Therefore, we further exclude the studies that did not report respective search results from the individual search sources, and then use the remaining publication set to calculate the rate of reproducible searches. The results are shown in Table 4(c).

Figure 2(c) shows the number of search sources is a critical factor that significantly affects the reproducibility of automated search. By contrast, venue type, once again, appears to have little effect. We observed that depending on how the researchers implemented and documented their secondary studies, even journal papers may lack clear search strategy descriptions, while the automated search reported in a short conference paper (e.g., [38]) can still be almost reproducible.

F. OVERALL CONCLUSIONS

Overall, we conclude that using fewer numbers of search sources increases the reproducibility of automated search in SE secondary studies. This conclusion does not deny the previous lessons about using multiple sources to enlarge the overage of primary studies [11]. Instead, we argue that it may not be wise to use many search sources in one study. According to the usage statistics (see Figure 5), the most practical trade-off seems to be employing four to six search sources (i.e., digital libraries and/or indexing platforms).

VI. GUIDELINES FOR REPRODUCIBLE SEARCH

Given the results reported in Section V, as well as our experiences of re-performing the searches, we formulate a set of evidence-based guidelines to help researchers ensure they formulate, execute and document reproducible searches. The guidelines are presented in Table 5.

We choose not to complement the guidelines with discussion and elaboration. We do this for three reasons: first, for the purpose of conciseness; second, and more significantly, because the guidelines are intended to be atomic and therefore self-contained; and third, by extension, we want the guidelines to be accessible. Furthermore, we focus on technical-level, specific guidelines rather than generic advice (i.e., we avoid recommendations such as, "Ensure the decision is justified," or "Ensure, you report X.") or advice already provided by others, such as Ali and Usman [13].

We do however want to emphasise the *formulation* of searches. If search strings are fundamentally flawed, e.g., they contain syntax errors, then no amount of transparency when documenting the searches can fix the original *results* of the search. More generally, Figure 1 and Figure 3 show that many problems with searches can potentially be addressed through better preparation, either of the string itself, of the related settings, or of the choice of search sources.

VII. THREATS TO VALIDITY

In this section we first consider potential threats to our evaluation. We then consider the implications of these threats for our guidelines.

A. POTENTIAL BIAS IN THE SAMPLING

We used a carefully selected set of 24 tertiary studies to identify a large sample (>1300) of secondary studies, and then selected 621 studies as our sample. Whilst our sample is relatively large, size does not in itself determine representativeness. As noted earlier, the use of peer-reviewed tertiary studies to identify a set of candidate secondary studies, and then the selection of peer-reviewed secondary studies from that candidate set, suggests we have a higher quality sample

TABLE 5. Guidelines for formulating, executing and documenting searches.

[Preparing for the searches]
1 Formulating search strings:
  a Ensure that search terms and the rules for combining them are decided and justified.
  b Use shorter strings to improve string reusability.
  c Do not rely on the search sources' default Boolean operations.
  d Explicitly specify Boolean operators between search terms.
  e Enter all Boolean operators (e.g., `AND` and `OR`) in capital case, not lowercase letter (e.g., not `And/Or` or `and/or`).
  e Do not use other logic symbols (e.g., $\wedge$, $\vee$) as Boolean operators.
  f To group search terms, use parentheses ( ) rather than other bracket types { }, [ ] or <>.
  g Ensure every multi-word search term is enclosed within a pair of quotation marks for exact matches of that phrase.
  h Ensure every hyphenated search term is enclosed within a pair of quotation marks for exact matches of that term.
  i Ideally, use quotation marks to enclose every term of a search string, even if the search term has a single word only.
  j Use straight quotes (" ") instead of curly quotes (“ ”) for exact matches of hyphenated words or multi-word phrases.
  k Ensure all quotation marks and parentheses are all paired correctly.
2 Selecting sources for the search:
  a WoS is recommended to help prepare and verify search strings, even if it is not selected as a search source.
  b Ensure you justify the selection of your search sources, e.g., see Figure 3.
  c Limited amount of search sources are preferred for improving the reproducibility of search.
  d Ensure the prepared search string does not violate the string length limit of the respective search source:
    i DBLP has a maximum limit of 127 characters; Google Scholar has a maximum limit of 256 characters.
    ii IGI Global has a maximum limit of 100 characters; JSTOR has a maximum limit of 200 characters.
    iii ScienceDirect has a maximum limit of 500 characters.
    iv SpringerLink does not work well with lengthy search strings (about 1800 characters).
  e Ensure the prepared search string does not violate the Boolean operator limit for the respective search sources:
    i ACM DL does not support the proximity operators.
    ii Google Scholar does not recognise wildcards and the Boolean operator `AND` in its Title search.
    iii IEEE Xplore has a maximum limit of seven wildcards in one string; ScienceDirect does not support wildcard search.
    iv ScienceDirect has a maximum limit of eight Boolean operators per search field.
    v WoS supports left-hand wildcard only for the Topic, Title, and Identifying Code searches.
  f Be cautious when using the following sources:
    i AIS eLibrary is not recommended for complex search strings, due to its bug of parentheses interpretation.
    ii CiteSeerX is not recommended in general, due to its unexplainable search behaviours.
    iii DBLP is not recommended in general, due to its unique query syntax and the lack of advanced search.
3 Using zonal settings:
  a Do not violate the field limit of the selected search source:
    i Google Scholar and SpringerLink do not support zonal settings, except for the Title search.
    ii ScienceDirect does not support exclusive abstract search or keywords search.
    iii ScienceDirect, SpringerLink and Wiley OL do not support command line search.
    iv SpringerLink's title search is not recommended, as it does not work as expected.
    v Wiley does not support disjunction between multiple search fields.
4 Using time range.
[Executing the searches]
5 Minimise adaptions to the search string. This requires a well-prepared generic string.
6 Use WoS to test the prepared string prior to use.
7 When searching on IEEE Xplore:
  a If the search field in the zonal setting is All Metadata only, use the default Command Search to avoid prefixing the field name to search terms.
  b If there are unavoidable parentheses in the search string, prefix the field name to every search term of the string.
[Documenting the searches]
8 Documenting Search Strings:
  a Report the truly used full search string(s)
  b For each engine-specific adaptation of the general search string, list the alternative strings used.
  c For LaTeX manuscripts, using a pair of `\textquotedbl` (instead of quotation marks from the keyboard) to output straight quotes.
  d For Word manuscripts, uncheck the option `"Straight quotes" with “smart quotes”` under both the `AutoFormat` tab and the `AutoFormat As You Type` tab in the `AutoCorrect` dialog box, for being able to output straight quotes.
  e Ensure the search string(s) is documented and reported in plain text instead of in a figure.
9 Documenting search sources:
  a Report the web addresses for the search sources.
10 Documenting time and zonal settings:
  a Report the truly used zonal settings and time range.
  b Report the date of the execution of the search.
11 Documenting search results:
  a Report the number of hits from each of the used search source.

(at least relatively). One implication is that our overall evaluation, e.g., Figure 3, may be “optimistic”, applying to the “better” secondary studies in SE.

A second threat with the sample is that while the overall sample is relatively large, not all of the sample could be used for the factorial ANOVAs, e.g., due to missing data.

Our detailed quantitative analyses may therefore suffer from sample size (see sample sizes in Table 4). An obvious way to address this threat is to increase the sample size, e.g., we might have used more of the approximate 1300 candidates. But increasing sample size may not improve data quality, e.g., may not address the problem of missing data.

B. POTENTIAL BIAS FROM NOT ACCESSING SOURCES

At the time we conducted our searches, the authors undertaking the searches had no access to Engineering Village (EV). We therefore could not test automated searches over it. But we still included EV in this research and assumed the corresponding tests to be passed, as long as the relevant data are available in the publications. For example, we have considered the automated search in [39] to be reproducible without physically testing it, because this study used EV as the only search source, and it reported all the critical data ranging from the search string to the search results. Nevertheless, the lack of access to EV could bias both our overall results (see Figure 1) and our factorial ANOVAs. We decided not to exclude Engineering Village, because it is a well-known indexing platform of engineering literature, and excluding it will lead to bigger bias in the usage statistics of search sources (see Section V-B). Retaining EV does not have significant impact on the conclusions of this research: the current results for the three indicators may be slightly higher than they should be, while the reproducibility crisis of automated search is still clearly revealed.

C. POTENTIAL BIAS IN THE FACTORIAL ANOVA

For the ANOVA, we had to split each factor's values into two groups to satisfy the requirement of two-level factorial ANOVA. Such a dichotomisation inevitably results in loss of information and usually involves subjective decisions. We used median averages to distinguish between few and many search term-counts and between a few and many search sources. We choose median averages because it is generally the best representative of a dataset's central location, especially when the data set has a skewed distribution [40]. Statistical conclusions could, of course, change if we replace median with the other measures of central tendency, such as mean or mode.

Also, our suggestions (i.e., using fewer search terms and less search sources to improve the reproducibility of automated search) are essentially driven by data correlation instead of causation. More importantly, these suggestions have nothing to do with the completeness of the relevant study selection. In other words, making automated search more reproducible does not guarantee a better coverage of relevant studies. However, it is also uncertain whether or not using more search terms and/or more search sources would guarantee an improvement in coverage. Furthermore, we argue that even if an irreproducible automated search brings 100% relevant studies, it will still harm the quality of the systematic review, as it will lead to doubts about whether and how the completeness is addressed.

D. IMPLICATIONS FOR THE GUIDELINES

Our guidelines are prescriptive for good practice rather than predictive of research quality. They are *derived* from our evaluation and from our experience of conducting the evaluation. As such, biases in our evaluation should have limited direct impact on the value of the guidelines. We recognise these guidelines need to be used and, through use, to be evaluated independently, and we encourage researchers to undertake such evaluation.

VIII. CONCLUSION AND FUTURE WORK

Automated search of search sources is increasingly common in SE research, the most obvious example being the systematic review. (Such searches may also be employed, albeit less formally, in other activities, such as the more traditional literature reviews.) Given the growth of systematic reviews in SE, and the potential impact of these reviews as a systematic summary of the state of evidence on a subject, we evaluated the reproducibility of automated searches. We focused our evaluation on secondary studies that are based on the systematic literature review protocol first proposed by Kitchenham and Charters [1]. We had two research objectives: first, to better understand whether searches were reproducible; second, to propose improvements to automated search.

For our first objective, we asked a research question, *viz.* to what extent can we reproduce automated searches from existing SE secondary studies? We distinguished reusable search strings from repeatable searches and from reproducible searches. We conclude that only about 6% of the 621 sample studies report reproducible searches; that about 13% report repeatable searches, and that about 47% report reusable search strings. Search is a fundamental stage for a secondary study, and search strings are a core element of such a stage. Our findings are concerning because they suggest very few secondary studies can be reproduced.

The outcome of our first research objective motivated our second objective, i.e., how can we improve the reproducibility of automated search in SE secondary studies? Our response to this question has been to propose a set of low-level, specific guidelines that are derived from our evaluation and from our experience of re-performing the searches. The guidelines concentrate on the formulation and execution of searches.

Overall, our article contributes an evaluation of the state-of-practice of automated searches in secondary studies in SE, and a set of guidelines for improving the reproducibility of future secondary studies.

Given the contributions of this article, we propose two directions for further research. First, to develop, apply and evaluate a tool, or similar, to facilitate search-string preparation and adaptation. This tool might be further extended to obtain candidate publications (or their metadata) by programmatically executing automated search in suitable cases, e.g., when APIs are available, or Web crawling is enabled.

Second, to investigate the "fitness for purpose" of a search string against its corresponding review's research question(s). We have evaluated whether a search is reproducible. We also want to investigate the efficacy of a search.

REFERENCES

[1] B. A. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," School Comput. Sci. Math., Keele Univ. Durham Univ., Tech. Rep. EBSE 2007-01, Jul. 2007.

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

[3] S. Barat, T. Clark, B. Barn, and V. Kulkarni, "A model-based approach to systematic review of research literature," in *Proc. 10th Innov. Softw. Eng. Conf.*, Feb. 2017, pp. 15–25.

[4] X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang, "A map of threats to validity of systematic literature reviews in software engineering," in *Proc. 23rd Asia–Pacific Softw. Eng. Conf. (APSEC)*, Dec. 2016, pp. 153–160.

[5] Y. Shakeel, J. Krüger, I. von Nostitz-Wallwitz, C. Lausberger, G. C. Durand, G. Saake, and T. Leich, "(Automated) literature analysis—Threats and experiences," in *Proc. IEEE/ACM 13th Int. Workshop Softw. Eng. Sci. (SE4Science)*, Jun. 2018, pp. 20–27.

[6] S. Imtiaz, M. Bano, N. Ikram, and M. Niazi, "A tertiary study: Experiences of conducting systematic literature reviews in software engineering," in *Proc. 17th Int. Conf. Eval. Assessment Softw. Eng.*, Apr. 2013, pp. 177–182.

[7] J. Bailey, C. Zhang, D. Budgen, M. Turner, and S. Charters, "Search engine overlaps: Do they agree or disagree?" in *Proc. 2nd Int. Workshop Realising Evidence-Based Softw. Eng. (REBSE)*, May 2007, pp. 1–6.

[8] Z. Li, "Stop building castles on a swamp! The crisis of reproducing automatic search in evidence-based software engineering," in *Proc. IEEE/ACM 43rd Int. Conf. Softw. Eng., New Ideas Emerg. Results (ICSE-NIER)*, May 2021, pp. 16–20.

[9] J. Krüger, C. Lausberger, I. von Nostitz-Wallwitz, G. Saake, and T. Leich, "Search. Review. Repeat? An empirical study of threats to replicating SLR searches," *Empirical Softw. Eng.*, vol. 25, no. 1, pp. 627–677, Jan. 2020.

[10] P. Singh and K. Singh, "Exploring automatic search in digital libraries: A caution guide for systematic reviewers," in *Proc. 21st Int. Conf. Eval. Assessment Softw. Eng.*, Jun. 2017, pp. 236–241.

[11] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, Apr. 2007.

[12] T. Dyba, T. Dingsoyr, and G. K. Hanssen, "Applying systematic reviews to diverse study types: An experience report," in *Proc. 1st Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Sep. 2007, pp. 225–234.

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

[14] D. C. Mariano, C. Leite, L. H. S. Santos, R. E. O. Rocha, and R. C. de Melo Minardi, "A guide to performing systematic literature reviews in bioinformatics," Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasília, Brazil, Tech. Rep. RT.DCC.002/2017, 2017.

[15] C. Okoli, "A guide to conducting a standalone systematic literature review," *Commun. Assoc. Inf. Syst.*, vol. 37, p. 43, Nov. 2015.

[16] *Reproducibility and Replicability in Science*, National Academies of Sciences, Engineering, and Medicine, National Academies, Washington, DC, USA, 2019.

[17] R. L. Novais, A. Torres, T. S. Mendes, M. Mendonça, and N. Zazworka, "Software evolution visualization: A systematic mapping study," *Inf. Softw. Technol.*, vol. 55, no. 11, pp. 1860–1883, Nov. 2013.

[18] M. Galster, D. Weyns, D. Tofan, B. Michalik, and P. Avgeriou, "Variability in software systems—A systematic literature review," *IEEE Trans. Softw. Eng.*, vol. 40, no. 3, pp. 282–306, Mar. 2014.

[19] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng.*, May 2014, pp. 1–10.

[20] Z. Li, P. Avgeriou, and P. Liang, "A systematic mapping study on technical debt and its management," *J. Syst. Softw.*, vol. 101, pp. 193–220, Mar. 2015.

[21] B. J. Williams and J. C. Carver, "Characterizing software architecture changes: A systematic review," *Inf. Softw. Technol.*, vol. 52, no. 1, pp. 31–51, Jan. 2010.

[22] H. Zhang and M. A. Babar, "On searching relevant studies in software engineering," in *Proc. Electron. Workshops Comput.*, Apr. 2010, pp. 111–120.

[23] E. Bayram, B. Dogan, and V. Tunali, "Bibliometric analysis of the tertiary study on agile software development using social network analysis," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Oct. 2020, pp. 1–4.

[24] R. E. Lopez-Herrejon, L. Linsbauer, and A. Egyed, "A systematic mapping study of search-based software engineering for software product lines," *Inf. Softw. Technol.*, vol. 61, pp. 33–51, May 2015.

[25] N. Salleh, E. Mendes, and J. Grundy, "Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review," *IEEE Trans. Softw. Eng.*, vol. 37, no. 4, pp. 509–525, Jul. 2011.

[26] V. Garousi and B. Küçük, "Smells in software test code: A survey of knowledge in industry and academia," *J. Syst. Softw.*, vol. 138, pp. 52–81, Apr. 2018.

[27] D. C. Montgomery, *Design and Analysis of Experiments*, 9th ed. Hoboken, NJ, USA: Wiley, Apr. 2019.

[28] R. D. S. Rocha and M. Fantinato, "The use of software product lines for business process management: A systematic literature review," *Inf. Softw. Technol.*, vol. 55, no. 8, pp. 1355–1373, Aug. 2013.

[29] Z. Li and A. Rainer, "Academic search engines: Constraints, bugs, and recommendations," in *Proc. 13th Int. Workshop Automating Test Case Design, Selection Eval.*, Nov. 2022, pp. 25–32.

[30] D. M. Fernández, S. Ognawala, S. Wagner, and M. Daneva, "Where do we stand in requirements engineering improvement today: First results from a mapping study," in *Proc. 8th ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas.*, Sep. 2014, pp. 18–19.

[31] 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. Softw.*, vol. 86, no. 10, pp. 2594–2610, Oct. 2013.

[32] M. Hosseini, A. Shahri, K. Phalp, J. Taylor, and R. Ali, "Crowdsourcing: A taxonomy and systematic mapping study," *Comput. Sci. Rev.*, vol. 17, pp. 43–69, Aug. 2015.

[33] K. A. Alam, R. Ahmad, A. Akhunzada, M. H. N. M. Nasir, and S. U. Khan, "Impact analysis and change propagation in service-oriented enterprises: A systematic review," *Inf. Syst.*, vol. 54, pp. 43–73, Dec. 2015.

[34] M. Shahin, P. Liang, and M. A. Babar, "A systematic review of software architecture visualization techniques," *J. Syst. Softw.*, vol. 94, pp. 161–185, Aug. 2014.

[35] P. H. Nguyen, M. Kramer, J. Klein, and Y. L. Traon, "An extensive systematic review on the model-driven development of secure systems," *Inf. Softw. Technol.*, vol. 68, pp. 62–81, Dec. 2015.

[36] K. Wnuk and R. K. Kollu, "A systematic mapping study on requirements scoping," in *Proc. 20th Int. Conf. Eval. Assessment Softw. Eng.*, Jun. 2016, pp. 1–3.

[37] L. Chen, M. A. Babar, and H. Zhang, "Towards an evidence-based understanding of electronic data sources," in *Proc. Electron. Workshops Comput.*, Apr. 2010, pp. 135–138.

[38] M. Anjum and D. Budgen, "A mapping study of the definitions used for service oriented architecture," in *Proc. 16th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, May 2012, pp. 57–61.

[39] Z. Sharafi, Z. Soh, and Y.-G. Guéhéneuc, "A systematic literature review on the usage of eye-tracking in software engineering," *Inf. Softw. Technol.*, vol. 67, pp. 79–107, Nov. 2015.

[40] J. Frost. (Feb. 12, 2018). *Measures of Central Tendency: Mean, Median, and Mode*. [Online]. Available: https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

SELECTED TERTIARY STUDIES

[T1] J. L. Barros-Justo, F. B. V. Benitti, and S. Matalonga, "Trends in software reuse research: A tertiary study," *Comput. Standards Interface*, vol. 66, Oct. 2019, Art. no. 103352.

[T2] D. Budgen, P. Brereton, S. Drummond, and N. Williams, "Reporting systematic reviews: Some lessons from a tertiary study," *Inf. Softw. Technol.*, vol. 95, pp. 62–74, Mar. 2018.

[T3] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli, ''Usability in agile software development: A tertiary study,'' *Comput. Standards Interface*, vol. 64, pp. 61–77, May 2019.

[T4] M. Goulão, V. Amaral, and M. Mernik, ''Quality in model-driven engineering: A tertiary study,'' *Softw. Quality J.*, vol. 24, no. 3, pp. 601–633, Sep. 2016.

[T5] R. Hoda, N. Salleh, J. Grundy, and H. M. Tee, ''Systematic literature reviews in agile software development: A tertiary study,'' *Inf. Softw. Technol.*, vol. 85, pp. 60–70, May 2017.

[T6] A. A. Khan, J. Keung, M. Niazi, S. Hussain, and H. Zhang, ''Systematic literature reviews of software process improvement: A tertiary study,'' in *Proc. 24th Eur. Conf. Softw. Process Improvement (EuroSPI)*. Ostrava, Czech Republic: Springer, Sep. 2017, pp. 177–190.

[T7] C. Marimuthu and K. Chandrasekaran, ''Systematic studies in software product lines: A tertiary study,'' in *Proc. 21st Int. Syst. Softw. Product Line Conf. A*, Sep. 2017, pp. 143–152.

[T8] I. Nurdiani, J. Börstler, and S. A. Fricker, ''The impacts of agile and lean practices on project constraints: A tertiary study,'' *J. Syst. Softw.*, vol. 119, pp. 162–183, Sep. 2016.

[T9] M. Raatikainen, J. Tiihonen, and T. Männistö, ''Software product lines and variability modeling: A tertiary study,'' *J. Syst. Softw.*, vol. 149, pp. 485–510, Mar. 2019.

[T10] N. Rios, M. G. D. Mendonça Neto, and R. O. Spínola, ''A tertiary study on technical debt: Types, management strategies, research trends, and base information for practitioners,'' *Inf. Softw. Technol.*, vol. 102, pp. 117–145, Oct. 2018.

[T11] Y. Zhou, H. Zhang, X. Huang, S. Yang, M. A. Babar, and H. Tang, ''Quality assessment of systematic reviews in software engineering: A tertiary study,'' in *Proc. 19th Int. Conf. Eval. Assessment Softw. Eng.*, Apr. 2015, pp. 1–14.

[T12] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, ''Identifying, categorizing and mitigating threats to validity in software engineering secondary studies,'' *Inf. Softw. Technol.*, vol. 106, pp. 201–230, Feb. 2019.

[T13] Z. I. Batouta, R. Dehbi, M. Talea, and O. Hajoui, ''Automation in code generation: Tertiary and systematic mapping review,'' in *Proc. 4th IEEE Int. Colloq. Inf. Sci. Technol. (CiSt)*, Oct. 2016, pp. 200–205.

[T14] D. Budgen, P. Brereton, N. Williams, and S. Drummond, ''The contribution that empirical studies performed in industry make to the findings of systematic reviews: A tertiary study,'' *Inf. Softw. Technol.*, vol. 94, pp. 234–244, Feb. 2018.

[T15] H. Cadavid, V. Andrikopoulos, and P. Avgeriou, ''Architecting systems of systems: A tertiary study,'' *Inf. Softw. Technol.*, vol. 118, Feb. 2020, Art. no. 106202.

[T16] B. Cartaxo, G. Pinto, D. Ribeiro, F. Kamei, R. E. S. Santos, F. Q. B. da Silva, and S. Soares, ''Using Q&A websites as a method for assessing systematic reviews,'' in *Proc. IEEE/ACM 14th Int. Conf. Mining Softw. Repositories (MSR)*, May 2017, pp. 238–242.

[T17] S. Elmidaoui, L. Cheikhi, and A. Idri, ''Software product maintainability prediction: A survey of secondary studies,'' in *Proc. 4th Int. Conf. Control, Decis. Inf. Technol. (CoDIT)*, Apr. 2017, pp. 702–707.

[T18] C. Fu, H. Zhang, X. Huang, X. Zhou, and Z. Li, ''A review of meta-ethnographies in software engineering,'' in *Proc. Eval. Assessment Softw. Eng.*, Apr. 2019, pp. 68–77.

[T19] V. Garousi and M. V. Mäntylä, ''A systematic literature review of literature reviews in software testing,'' *Inf. Softw. Technol.*, vol. 80, pp. 195–216, Dec. 2016.

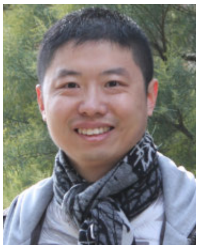
[T20] H. G. Gürbüz and B. Tekinerdogan, ''Software metrics for green parallel computing of big data systems,'' in *Proc. IEEE Int. Congr. Big Data (BigData Congress)*, Jun. 2016, pp. 345–348.

[T21] A. Idri and L. Cheikhi, ''A survey of secondary studies in software process improvement,'' in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2016, pp. 1–8.

[T22] M. U. Khan, S. Sherin, M. Z. Iqbal, and R. Zahid, ''Landscaping systematic mapping studies in software engineering: A tertiary study,'' *J. Syst. Softw.*, vol. 149, pp. 396–436, Mar. 2019.

[T23] G. Lacerda, F. Petrillo, M. Pimenta, and Y. G. Guéhéneuc, ''Code smells and refactoring: A tertiary systematic review of challenges and observations,'' *J. Syst. Softw.*, vol. 167, Sep. 2020, Art. no. 110610.

[T24] A. Yasin, R. Fatima, L. Wen, W. Afzal, M. Azhar, and R. Torkar, ''On using grey literature and Google scholar in systematic literature reviews in software engineering,'' *IEEE Access*, vol. 8, pp. 36226–36243, 2020.



ZHENG LI (Senior Member, IEEE) received the B.Eng. degree from Zhengzhou University, the M.Sc.Eng. degree from the Beijing University of Chemical Technology, the M.Phil. degree from the University of New South Wales (UNSW), and the Ph.D. degree from Australian National University (ANU). During the same time, he was a Graduate Researcher with the Software Systems Research Group (SSRG), National ICT Australia (NICTA). Before studying abroad, he had around four years of industrial experience in China. He is currently a Lecturer with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, U.K. Previously, he was a tenured Assistant Professor with the Department of Computer Science, University of Concepción, Chile. Before that, he was a Postdoctoral Researcher with the Cloud Control Group, Lund University, Sweden. He was also a Visiting Research Fellow with the Software Institute, Nanjing University, China.

AUSTEN RAINER is currently a Professor with the School of Electronics, Electrical Engineering and Computer Science (EEECS), Queen's University Belfast (QUB), where he is the Theme Lead of the School's Software Engineering Strategic Research Theme, which he established in 2020. He is also a Mentor of the Queen's Gender Initiative, QUB. Previously, he was the Head of the Department of Computer Science and Software Engineering, University of Canterbury, New Zealand. Whilst in New Zealand, he co-founded and become the Founding Chair of Software Innovation NZ, a national software researchers' network. He was a member of the Establishment Team for the successful formation of a five-institute South Island ICT Graduate School, the first of its kind in New Zealand. He was a member of the Multidisciplinary MOOD Movement Network (ESRC; ES/V00848X/1). For six months, he was a Technical Advisor to a technology start-up in Northern Ireland that successfully received several rounds of funding, including a fellowship from the RAE. Since 2020, he has co-facilitated positive academic leadership development courses for Ph.D. students through Informatics Europe.