e-Informatica Software Engineering Journal, Volume 16, Issue 1, 2022, pages: 220106, DOI: 10.37190/e-Inf220106

A Comparison of Citation Sources for Reference and Citation-Based Search in Systematic Literature Reviews

Nauman bin Ali*, Binish Tanveer*
*Blekinge Institute of Technology, Sweden
nauman.ali@bth.se, binish.tanveer@bth.se

Abstract

Context: In software engineering, snowball sampling has been used as a supplementary and primary search strategy. The current guidelines recommend using Google Scholar (GS) for snowball sampling. However, the use of GS presents several challenges when using it as a source for citations and references.
Objective: To compare the effectiveness and usefulness of two leading citation databases (GS and Scopus) for use in snowball sampling search.
Method: We relied on a published study that has used snowball sampling as a search strategy and GS as the citation source. We used its primary studies to compute precision and recall for Scopus.
Results: In this particular case, Scopus was highly effective with 95% recall and had better precision of 5.1% compared to GS's 2.8%. Moreover, Scopus found nine additional relevant papers. On average, one would read approximately 15 extra papers in GS than Scopus to identify one additional relevant paper. Furthermore, Scopus supports batch downloading of both citations and papers' references, has better quality metadata, and does better source filtering.
Conclusion: This study suggests that Scopus seems to be more effective and useful for snowball sampling than GS for systematic secondary studies attempting to identify peer-reviewed literature.

Keywords: Snowball sampling, snowballing, reference-based, citation-based, search strategy, systematic review, systematic mapping

1. Introduction

Systematic literature reviews and mapping studies [1] rely on a systematic and extensive search to identify the literature on a topic of interest. The two main search strategies in such secondary studies have been: (1) the use of keyword-based search and (2) supplementing the keyword-based results with snowball sampling [2]. However, others have proposed to use snowball sampling as the primary search method [3, 4]. Snowball sampling refers to the use of reference-of (for backward snowballing) and citations-to (for forward snowballing), a set of papers for identifying other relevant papers.

The indexing/citation database plays a critical role, whether using snowball sampling as the primary or supplementary search strategy. The coverage of the citation database may limit the snowball sampling strategy's effectiveness. Several alternative electronic data

Submitted: 28 Sep. 2021; Revised: 08 Mar. 2022; Accepted: 09 Mar. 2022; Available online: 17 Mar. 2022

sources for citation search exist, e.g., Elsevier Scopus (Scopus), Clarivate Analytics – Web of Science (WoS), and Google Scholar (GS). However, the current guidelines [4] recommend using GS.

For keyword-based search, where an automatic search is conducted using a combination of keywords, several studies have investigated the relevance and coverage of different electronic data sources [5–7]. However, no such investigation is reported of electronic data sources for snowball sampling in software engineering (SE) to the best of our knowledge.

In this study, we have compared Scopus and GS for use in the snowball sampling search strategy. We choose GS and Scopus, as these are among the most used citation databases in SE systematic reviews [2]. Similarly, the snowball sampling guidelines in SE recommend the use of GS [4]. The snowball sampling guidelines [4] already have over 2000 citations[1]. At least 1150 of these 2000 citing articles mention "google scholar"[2] indicating that GS is one of the sources used in these articles. This further justifies this study as the results could potentially have significant implications for future SE research employing snowball sampling in their search strategy.

The remainder of the paper is structured as follows: Section 2 describes the related work. Section 3 presents the approach used in this study for comparing the two sources. Section 4 presents the results. In Section 5 we further discussed the limitations of GS in light of related research. Section 6 presents our recommendations for future studies using snowball sampling search strategy. Section 7 highlights the validity threats and limitations of the paper. Section 8 concludes the paper.

2. Related work

We discuss the related work for this study in three complementary themes: (1) guidelines for the design, reporting, and evaluation of search strategies (2) the evaluation of electronic data sources used in SE, and (3) studies comparing citation databases without a focus on SE.

2.1. Search guidelines for systematic secondary studies

Several comprehensive guidelines for designing keyword-based search [1] and snowball sampling [4] are available. Furthermore, new improvements have been suggested to the design and assessment guidelines [8–11] based on the limitations identified in the repeatability of search in existing SLRs [2, 9].

Even when using keyword-based search as the primary search method, it is recommended to supplement the search using snowball sampling [1]. Thus, both search strategies will benefit from this study that assesses the usefulness and effectiveness of the currently recommended citation source.

2.2. Comparison of databases for keyword-based search in SE

There have been numerous studies investigating electronic data sources for keyword-based searches covering topics such as features required to support secondary studies ([7, 12, 13]) overlaps among sources ([5, 6, 14, 15]), and the value of Google and GS ([7, 13]). These

[1] On September 20, 2021, in GS, Wohlin's guidelines [4] had accumulated over 2000 citations.

[2] In GS, we searched for "Google Scholar" within the articles citing Wohlin's guidelines [4].

studies conclude that: (a.) multiple sources should be searched ([5, 6, 14, 15]), and (b.) GS and Google have good coverage of SE literature and SE secondary studies ([7, 13]).

However, we could not find a study in SE that compared citation databases for snowball sampling. In this study, we fill this gap by assessing the effectiveness and usefulness of the recommended citation database, i.e., GS, in the current guidelines and comparing it with other commonly used citation databases in SE research.

2.3. Studies comparing citation databases outside SE

Several comparisons of citations databases have been conducted, which are summarised in Table 1. The main sources compared include GS, WoS, Scopus, and Microsoft Academic Search (MAS). Some studies have indicated that the coverage (both in terms of indexing papers and citations) varies between different sources depending on the timeframe and research areas considered [16, 17]. However, one can conclude that overall, GS has the most indexed bibliographic records and citations [18–21]. GS also seems to have a faster indexing speed [22]. On the other hand, the quality of data and transparency of what is indexed is better in paid services like Scopus [22].

In this study, which is the first in SE literature, we compare GS and Scopus as sources for citations and references when performing snowball sampling. Unlike the studies discussed

Table 1. An overview of related work on comparing various data sources outside SE

ID	Data sources compared	Parameters	Main conclusions
[16]	GS, WoS	As a source for forward snowball sampling for public health literature.	WoS is recommended for public health guidance needs.
[17]	GS, WoS, Scopus	For citation tracking in two fields oncology and condensed matter physics.	Databases performance varied for different research areas and publication years.
[18]	GS, MAS	No of papers indexed and citation to those papers for several authors.	GS indexed more papers and citations for information and computing literature.
[22]	GS, Scopus	Indexed sources and indexing speed.	Scopus provides a clear documentation of what is indexed in its database. Scopus has higher accuracy and quality of data.The most important additional source indexed by GS is Google Books. GS has faster indexing speed.
[19]	GS, WoS, Scopus, MAS, and eight others	Number of bibliographic records indexed by the data source.	GS had the most number of bibliographic records.
[20]	GS, WoS, Scopus, MAS, and two others.	The number of citations to a set of 2515 highly-cited documents.	GS has the most number of citations.
[21]	GS, WoS, Scopus, MAS.	The number of citations to set of 150 articles from journals with high, low and no impact factors.	GS and Microsoft Academic had similar average number of citations, which were much higher than WoS and Scopus.

above, we take into consideration the relevance of the additional citations (not just the number of citations) found by a source.

3. Research method

In this study, we have only compared Scopus and GS. These two are the most often used [2] citation databases [6] in secondary studies in SE. Moreover, GS is the recommended source in snowball sampling guidelines in SE [4]. Hence, we attempt to answer the following research question:

RQ: *How effective and useful are GS and Scopus citation databases for implementing snowball sampling-based search strategy?*

To compare Scopus and GS, we have used a published systematic review [23] (from here on referred to as the case SLR) that has used GS for executing the snowball sampling search strategy. We did this for convenience as we had access to the intermediate and final results, which is hard to obtain for papers where one has not been a co-author [24]. In the future, we intend to replicate the analysis reported in this paper on more published papers reducing the bias that having a single case introduces. This limitation is further discussed in Section 7. The data used in the study is available for replication and further analysis by other researchers at this link.

3.1. Criteria for assessing the effectiveness and usefulness

We now present the criteria for evaluating the effectiveness and usefulness of citation databases as used in this study:

Effectiveness: The primary studies from the case SLR have allowed us to objectively assess the implication of using Scopus instead of GS using the following metrics (adapted from [1, 16]):

- **Recall** = *100 * (# of primary studies found in the search) / (total # of primary studies).*
- **Precision** = *100 * (total # of primary studies in the search results) / (total # of search results).*
- **Number needed to read for each relevant paper (NNR)** = *(total # of excluded papers) / (# of primary studies found in the search results)*

Usefulness: Usefulness [25] in this study is defined as a subjective measure of how well a source supports users performing snowball sampling. We consider the following features as indicators for the usefulness of a citation source for snowball sampling:

- Ability to easily download citations to a paper.
- Ability to easily download the references in a paper.
- Ability to easily filter citations and references (e.g., based on the publication language, venue, or whether they are peer-reviewed).

This is not an exhaustive list of features. However, these are essential for enabling the use of snowball sampling as a search approach.

3.2. Overview of the relevant aspects of the case SLR

The case SLR [23] was an attempt to find industrially relevant regression testing research. Existing SLRs on the topic of regression testing were identified and used as a start-set for one iteration of forward and backward snowball sampling. Forward snowballing here refers to reviewing the citations to the papers in the start-set, and backward snowballing refers to checking the references used in the papers in the start-set. By one-iteration, we mean that no further snowballing was performed on the additionally included relevant papers found in the first iteration.

The search in the case SLR [23] was done in August 2016. Since the search was done in August (without a clear cut-off at a full year), it was impossible to recreate the citations list of Scopus in August 2016 at the time of the current study. Therefore, to have a relatively fair comparison, we have included the citations from both GS and Scopus up to and including the calendar year 2016.

Table 2 provides information about the start set, the number of citations and references in the start set. The case SLR had 38 primary studies. However, four papers were excluded from the comparison in the current study (i.e., making 34 primary studies in Table 2). Of the four papers not considered as primary studies in the current review, three were excluded as these were not identified by snowball sampling in the case SLR (these were added as a known-set of papers), and the remaining one paper was in pre-print in 2016, i.e., at the time of the search in the case SLR. The paper was eventually printed in 2017. This means that a search now will not find it as a publication in 2016 but as a publication from 2017. Furthermore, we had identified 12 of these primary studies through backward snowballing (i.e., through references in the seed set and do not represent the value of the data source, since they are listed in the full-text of the papers). Therefore, when assessing the effectiveness of GS and Scopus, we have used only 22 primary studies found by forward snowballing for comparison. For assessing the usefulness, we consider the features of the citation sources for both forward and backward snowball sampling.

Table 2. Details of the start set used in the case SLR [23]

No. of papers in seed set	11
No. of references in the seed set	877
No. of unique references in the seed set	506
No. of primary studies	34
Primary studies found through backward snowballing only	12

4. Results

After removing duplicate citations, we had 764 citations in GS and 415 in Scopus (see Table 3). Table 3 shows the number of citations and the objective measures of effectiveness for both sources. Please note that recall by definition will be 100% for the source used as a baseline. We also present the potential impact of having used Scopus in the Table 3.

The Venn diagram (see Figure 1) shows that 365 papers (that did not meet the inclusion criteria of the case SLR) and 21 (primary studies) are shared between Scopus and GS. At the same time, Scopus and GS each have 20 and 377 unique papers (that is papers that did not meet the inclusion criteria of the case SLR), respectively. Whereas nine potential primary studies are identified only by Scopus, and GS only identifies one unique primary study.

Table 3. Precision and recall for the two sources

	GS	Scopus
No. of citations to the seed papers	937	498
Unique citations to the seed papers (after removing duplicates)	764	415
Using only GS for forward snowballing, and its comparison with Scopus shows the following:[a]		
Of the 22 primary studies identified in GS	22	21
Precision	(22/764) * 100 = 2.8%	(21/415) * 100 = 5.1%
Recall	(22/22) * 100 = 100%	(21/22) * 100 = 95%
NNR	(377 + 365)/22 = 33.7	(29 + 365)/21 = 18.8
Using only Scopus for forward snowballing, and its comparison with GS shows the following:[b]		
Of the 30 potential primary studies identified in Scopus	21	30
Precision	(21/764) * 100 = 2.7%	(30/415) * 100 = 7.2%
Recall	(21/30) * 100 = 70.0%	(30/30) * 100 = 100%
NNR	(377 + 365 + 1)/21 = 35.3	(20 + 365)/30 = 12.8

[a] The nine potentially relevant papers only identified by Scopus are not considered in the analysis.
[b] One primary study only identified by GS is not considered in the analysis.

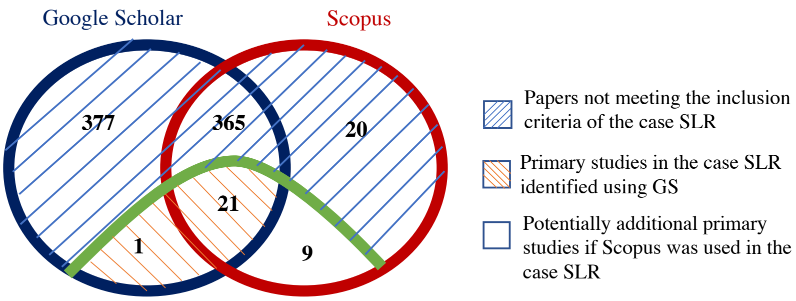
Google Scholar
Scopus
377
365
20
21
1
9
▨ Papers not meeting the inclusion criteria of the case SLR
▨ Primary studies in the case SLR identified using GS
□ Potentially additional primary studies if Scopus was used in the case SLR

Figure 1. Data and results of the comparison between GS and Scopus

4.1. Effectiveness of GS and Scopus

Of the 22 primary studies identified through forward snowballing in GS, except for one paper, we found all the primary studies through Scopus as well (see Figure 1). While both the missing primary study and the referenced seed paper are indexed in Scopus, the citation is not recognized in Scopus. We reported the issue to Elsevier's support, and the papers are now correctly linked.

We further analysed the 29 unique citations in Scopus (see Figure 1) by applying the selection criteria of the case SLR. We found that nine of these papers meet the selection criteria and would be shortlisted for data extraction and synthesis. However, we have not done the data extraction and re-analysis of the entire data for the current paper as we do not consider it essential for the objective of this paper. The impact of these nine papers on the metrics of effectiveness used in the study is presented in Table 3. The numbers indicate that Scopus would have been a far superior choice. However, since we have not done data extraction from these nine potential primary studies, we are not confident if they will all become primary studies. Therefore, for the remainder of the paper, we will focus on the numbers based on the case SLR where GS was used.

For the case SLR, the values for precision and NNR show that Scopus is more effective than GS. Scopus found 95% (21 out of the 22 papers) of the relevant papers identified by GS with considerably higher precision. The NNR value in (see Table 3) suggests that, on average, one would have to examine 15 extra papers in GS than Scopus to identify an additional relevant paper.

4.2. Usefulness of GS and Scopus

Table 4 summarises our assessment of GS and Scopus against the stated criteria for usefulness.

Table 4. Usefulness of GS and Scopus for snowball sampling

	GS	Scopus
Ability to easily download citations to a paper	No	Yes
Ability to easily download the references in a paper	No	Yes
Ability to easily filter citations and references	No	Partially yes[a]

[a] In Scopus, of the 14 fields of metadata to use for filtering citations only four fields are available to filter references in a paper. This was last confirmed in December 2021.

Downloading citations to a paper and references in a paper: In GS, it is difficult to download citations to papers. There is no native support for batch downloading of citations. Furthermore, to prevent denial of service attacks, GS blocks any attempt to automate the download. For example, one of the seed papers for the case SLR has over 1200 citations making it very difficult to download the citations manually. Furthermore, GS has no support for backward snowballing as references in the papers have to be manually extracted from the papers' full text.

On the other hand, we found that Scopus facilitates both forward and backward snowballing, by enabling batch download of citations and references.

Filtering citations and references: In GS, we found no means to exclude based on the publication language or whether they have been peer reviewed. For systematic studies that only include peer-reviewed literature published in certain languages (which is often the case in SE), we consider this a significant limitation of GS. Furthermore, due to the quality of metadata in GS, it was also difficult to remove duplicates. It took considerable effort to resolve minor differences in the titles and venues of the papers.

In Scopus, we can extract additional metadata about the publication, including the publication type and language that significantly aids in the selection process. Moreover, removing duplicates was reasonably straightforward in the citations retrieved through Scopus, as the data were relatively clean.

5. Discussion of GS in light of the related work

As discussed in Section 4, our study shows that GS does not have features that are necessary for its use in snowball sampling-based search. Furthermore, we found that Scopus was more effective and useful for this purpose. To further strengthen our recommendation to use Scopus instead of GS, we now briefly discuss the limitations of GS in terms of the nature and quality of metadata indexed in it, transparency of what is indexed, and its support for snowball sampling. We base this section both on the results of our study and also on investigations of GS by others.

5.1. Lack of transparency in what is indexed

There is a lack of transparency regarding what is indexed in GS [16, 26] which may explain to a certain degree the changing citation numbers for the same period [27]. This is a serious threat to the reliability of search when using GS. Furthermore, Winter et al. [27] concluded in a longitudinal study that the number of citations substantially increased in GS for the same articles retroactively (i.e., when the search was repeated for the number of citations for a paper in the same time period on a later date, a larger number of citations was retrieved). They conclude that coverage seems to have stabilized over the more recent years [27]. However, in a recent investigation Martín-Martín and López-Cózar [26] found large fluctuations in coverage of literature by GS, which they conclude is a clear limitation of GS's use as a data source for bibliometrics.

5.2. Quality of metadata

GS does not facilitate automatic data collection (see Section 4), and researchers use custom web scrapers to extract the list of citing documents (e.g., see Martín-Mart í [20]). For the current study, we used Publish or Perish[3]. However, we noticed several shortcomings in the collected data, e.g., several entries were missing venues, abstract, or publication years. This is consistent with the observations by other researchers [22, 28–30]. For example, Adriaanse and Rensleigh [30] compared the content quality of WoS, GS, and Scopus and found that Scopus outperformed both WoS and GS [30]. They concluded that GS had the most inconsistencies, like mistakes in author spellings and order and the volume and issue numbers for the publications. Recent bibliometric studies using citation data in various disciplines including SE have also used Scopus [31, 32].

5.3. The quality of literature in GS

Aguillo [33] investigated the literature coverage by GS by analyzing which web domains are the sources for their records. GS indexes low-quality literature like low-impact journals, teaching material, unpublished reports. They concluded that GS lacks the quality to use in bibliometric studies, a conclusion shared by other studies [22].

GS may be a useful source for studies interested in both peer-reviewed and non-peer-reviewed literature, e.g., in multi-vocal literature reviews [34] or topics wherein insufficient scientific literature is available.

However, there is considerable and unavoidable noise in GS search results for other studies, where primarily peer-reviewed literature is of interest. For example, the citations analysis of a paper with 234 citations in GS [35] revealed that only 116 of the 234 citations were from journal and conference papers in English, and 54 of the remaining 118 citations were from Grey-Literature.

[3] Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish.

6. Recommendations when using snowball sampling

The studies using snowball sampling as the search strategy often conflate a systematic literature study's search and selection phase. We have observed at least two consequences of this: (1) the level of record-keeping is insufficient for cross-validation and replications (in particular for studies considering a large number of papers), (2) it is challenging to employ the best practices for study selection (e.g., using multiple reviewers or using text mining-based solutions).

SLR authors need to record the meta information for each citation and reference considered in various snowball iterations in an SLR. Another benefit of documenting the start set, the metadata of papers considered and the finally included primary studies list will be to enable comparison of various citation sources for snowball sampling. However, several current SLRs using snowball sampling as the primary search strategy do not document the data about intermediate references and citations considered in an SLR.

Suppose Scopus is used to operationalize the snowballing strategy. Then with some additional effort, one can automatically download the citations and references and other necessary metadata, including publication venues, language, abstracts, and keywords. Once these references and citations are collated, and duplicates are removed (as done in the keyword-based search), we can proceed with using state-of-the-art procedures, and tools [24, 36, 37]) to assist the selection process [1].

Furthermore, the metrics and indicators used in this study [1, 6, 16] can be used to assess the electronic data sources for snowball sampling. However, these metrics must be interpreted in relative terms, i.e., to compare two or more data sources, as the entire population of all primary studies is unknown, and we typically only identify a subset of the primary studies in our search [38].

7. Validity threats

The current study has used only one case SLR; therefore, we need similar comparative analysis of other secondary studies to gain more confidence in the value of using Scopus. However, the results of the study illustrate the need to evaluate the recommendation of using GS in the guidelines for performing snowball sampling.

In this study, surprisingly (as several studies as discussed in Section 2 considered GS more comprehensive than Scopus), we found that Scopus has 29 unique contributions that are not available in GS. After applying the inclusion-exclusion criteria from the case SLR, we identified that nine of these papers would have been included in the case SLR. It will be interesting to see what may have been the impact of these on the results of the case SLR. However, that analysis has not been done in the current study since we did not consider it essential for the objective of this paper.

Since we are doing the study in 2021 and looking at citations in 2016, this may be a disadvantage to the database that is more efficient in indexing new publications and updating the citations. This limitation of our study can be overcome by replicating the analysis on more recently concluded SLRs that have used GS for snowball sampling.

We have cleaned the data extensively to avoid any problems, e.g., hyphenation or case differences in the citing papers' titles. However, we may have still missed a few unique cases where the same papers are considered unique due to slight differences. However, due to the measures taken and manual checking of some of the unique results, we are confident that this is not a significant threat to this study's validity.

Furthermore, the criteria used for usefulness are not very comprehensive. In the future, we should also collect additional data about the perceived usefulness of the two citation databases. However, we think that the criteria used for evaluation in this study indicate the usefulness of the databases for use in snowball sampling.

8. Conclusion

In this study, we have compared and empirically evaluated two leading alternative sources of citation data for snowball sampling. GS and Scopus have very different features and have different strengths, which will make them suitable for different use cases. However, based on the results of the current study, we conclude that Scopus is a superior source for snowball sampling in SE research when primarily peer-reviewed literature is targeted.

The results of this study suggest that by using Scopus instead of GS researchers can save substantial effort in data collection and reduce the effort spent on selection without a significant likelihood of missing relevant peer-reviewed literature. *Based on these findings, we recommend that the researchers employing a snowball sampling search strategy may use Scopus in the future.*

In the future, we would like to replicate the analysis reported in this study with other published secondary studies and with additional citation databases.

Acknowledgements

This work has been supported by ELLIIT, a Strategic Area within IT and Mobile Communications, funded by the Swedish Government and by research grants for the VITS project (reference number 20180127) and the SERT project from the Knowledge Foundation in Sweden.

References

[1] B.A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews.* Chapman & Hall/CRC, 2015.

[2] 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," *Empir. Softw. Eng.*, Vol. 25, No. 1, 2020, pp. 627–677.

[3] M. Skoglund and P. Runeson, "Reference-based search strategies in systematic reviews," in *13th International Conference on Evaluation and Assessment in Software Engineering, EASE*, Workshops in Computing, D. Budgen, M. Turner, and M. Niazi, Eds. Durham University, UK: BCS, 2009, pp. 31–40. [Online]. http://ewic.bcs.org/content/ConWebDoc/25022

[4] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *18th International Conference on Evaluation and Assessment in Software Engineering, EASE'14*, 2014, pp. 38:1–38:10.

[5] J. Bailey, C. Zhang, D. Budgen, M. Turner, and S. Charters, "Search engine overlaps: Do they agree or disagree?" in *2nd International Workshop on Realising Evidence-Based Software Engineering, REBSE'07*, 2007, p. 2.

[6] L. Chen, M.A. Babar, and H. Zhang, "Towards an evidence-based understanding of electronic data sources," in *14th International Conference on Evaluation and Assessment in Software Engineering, EASE.* BCS, 2010, pp. 135–138.

[7] A. Yasin, R. Fatima, L. Wen, W. Afzal, M. Azhar et al., "On using grey literature and Google Scholar in systematic literature reviews in software engineering," *IEEE Access*, Vol. 8, 2020, pp. 36 226–36 243.

[8] N. bin Ali and M. Usman, "A critical appraisal tool for systematic literature reviews in software engineering," *Inf. Softw. Technol.*, Vol. 112, 2019, pp. 48–50. [Online]. https://doi.org/10.1016/j.infsof.2019.04.006

[9] N. bin Ali and M. Usman, "Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy," *Information and Software Technology*, Vol. 99, 2018, pp. 133–147. [Online]. https://linkinghub.elsevier.com/retrieve/pii/S0950584917304263

[10] M. Usman, N. bin Ali, and C. Wohlin, "A quality assessment instrument for systematic literature reviews in software engineering," *CoRR*, Vol. abs/2109.10134, 2021. [Online]. https://arxiv.org/abs/2109.10134

[11] H.K.V. Tran, J. Börstler, N. bin Ali, and M. Unterkalmsteiner, "How good are my search strings? Reflections on using an existing review as a quasi-gold standard," *e-Informatica Software Engineering Journal*, Vol. 16, No. 1, 2022. [Online]. https://doi.org/10.37190/e-inf220103

[12] P. Singh and K. Singh, "Exploring automatic search in digital libraries: A caution guide for systematic reviewers," in *21st International Conference on Evaluation and Assessment in Software Engineering*, EASE'17. New York, NY, USA: ACM, 2017, pp. 236–241. [Online]. http://doi.acm.org/10.1145/3084226.3084275

[13] R. Fatima, A. Yasin, L. Liu, and J. Wang, "Google Scholar vs. dblp vs. Microsoft Academic Search: An indexing comparison for software engineering literature," in *44th Annual Computers, Software, and Applications Conference (COMPSAC)*. Madrid, Spain: IEEE, 2020, pp. 1097–1098. [Online]. https://ieeexplore.ieee.org/document/9202826/

[14] T. Dybå, T. Dingsøyr, and G.K. Hanssen, "Applying systematic reviews to diverse study types: An experience report," in *Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, ESEM*. ACM / IEEE Computer Society, 2007, pp. 225–234. [Online]. https://doi.org/10.1109/ESEM.2007.59

[15] J.A.M. Santos, A.R. Santos, and M.G. de Mendonça, "Investigating bias in the search phase of software engineering secondary studies," in *12th Workshop on Experimental Software Engineering*, 2015, pp. 488–501.

[16] P. Levay, N. Ainsworth, R. Kettle, and A. Morgan, "Identifying evidence for public health guidance: A comparison of citation searching with Web of Science and Google Scholar: Identifying Evidence for Public Health Guidance," *Research Synthesis Methods*, Vol. 7, No. 1, 2016, pp. 34–45.

[17] N. Bakkalbasi, K. Bauer, J. Glover, and L. Wang, "Three options for citation tracking: Google Scholar, Scopus and Web of Science," *Biomedical Digital Libraries*, Vol. 3, 2006.

[18] J. Ortega and I. Aguillo, "Microsoft Academic search and Google Scholar citations: Comparative analysis of author profiles," *Journal of the Association for Information Science and Technology*, Vol. 65, No. 6, 2014, pp. 1149–1156.

[19] M. Gusenbauer, "Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases," *Scientometrics*, Vol. 118, No. 1, 2019, pp. 177–214.

[20] A. Martín-Martín, M. Thelwall, E. Orduña-Malea, and E.D. López-Cózar, "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations," *Scientometrics*, Vol. 126, No. 1, 2021, pp. 871–906. [Online]. https://doi.org/10.1007/s11192-020-03690-4

[21] M. Levine-Clark and E. Gil, "A new comparative citation analysis: Google Scholar, Microsoft Academic, Scopus, and Web of Science," *Journal of Business and Finance Librarianship*, Vol. 26, No. 1–2, 2021, pp. 145–163.

[22] H.F. Moed, J. Bar-Ilan, and G. Halevi, "A new methodology for comparing Google Scholar and Scopus," *Journal of Informetrics*, Vol. 10, No. 2, 2016, pp. 533–551. [Online]. https://www.sciencedirect.com/science/article/pii/S1751157715302285

[23] N. bin Ali, E. Engström, M. Taromirad, M.R. Mousavi, N.M. Minhas et al., "On the search for industry-relevant regression testing research," *Empirical Software Engineering*, Vol. 24, No. 4, 2019, pp. 2020–2055.

[24] Z. Yu and T. Menzies, "FAST^2: An intelligent assistant for finding relevant papers," *Expert Syst. Appl.*, Vol. 120, 2019, pp. 57–71. [Online]. https://doi.org/10.1016/j.eswa.2018.11.021

[25] F.D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, 1989, pp. 319–340.

[26] A. Martín-Martín and E.D. López-Cózar, "Large coverage fluctuations in Google Scholar: A ase study," *CoRR*, Vol. abs/2102.07571, 2021. [Online]. https://arxiv.org/abs/2102.07571

[27] J.C.F.d. Winter, A.A. Zadpoor, and D. Dodou, "The expansion of Google Scholar versus Web of Science: A longitudinal study," *Scientometrics*, Vol. 98, No. 2, 2014, pp. 1547–1565.

[28] E.D. López-Cózar, E. Orduña-Malea, and A. Martín-Martín, "Google Scholar as a data source for research assessment," in *Springer Handbook of Science and Technology Indicators*, Springer Handbooks, W. Glänzel, H.F. Moed, U. Schmoch, and M. Thelwall, Eds. Springer, 2019, pp. 95–127. [Online]. https://doi.org/10.1007/978-3-030-02511-3_4

[29] G. Halevi, H. Moed, and J. Bar-Ilan, "Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation – Review of the literature," *Journal of Informetrics*, Vol. 11, No. 3, 2017, pp. 823–834.

[30] L. Adriaanse and C. Rensleigh, "Web of Science, Scopus and Google Scholar a content comprehensiveness comparison," *Electronic Library*, Vol. 31, No. 6, 2013, pp. 727–744.

[31] J.P. Ioannidis, K.W. Boyack, and J. Baas, "Updated science-wide author databases of standardized citation indicators," *PLoS Biology*, Vol. 18, No. 10, 2020, p. e3000918.

[32] K. Petersen and N. bin Ali, "An analysis of top author citations in software engineering and a comparison with other fields," *Scientometrics*, Vol. 126, No. 11, 2021, pp. 9147–9183. [Online]. https://doi.org/10.1007/s11192-021-04144-1

[33] I. Aguillo, "Is Google Scholar useful for bibliometrics? A webometric analysis," *Scientometrics*, Vol. 91, No. 2, 2012, pp. 343–351.

[34] V. Garousi, M. Felderer, and M.V. Mäntylä, "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," *Infiormation Software Technology*, Vol. 106, 2019, pp. 101–121. [Online]. https://doi.org/10.1016/j.infsof.2018.09.006

[35] N. bin Ali, H. Edison, and R. Torkar, "The impact of a proposal for innovation measurement in the software industry," in *ESEM'20: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, M.T. Baldassarre, F. Lanubile, M. Kalinowski, and F. Sarro, Eds. Bari, Italy: ACM, 2020, pp. 28:1–28:6. [Online]. https://doi.org/10.1145/3382494.3422163

[36] N. bin Ali and K. Petersen, "Evaluating strategies for study selection in systematic literature studies," in *ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM'14*, M. Morisio, T. Dybå, and M. Torchiano, Eds. Torino, Italy: ACM, 2014, pp. 45:1–45:4. [Online]. https://doi.org/10.1145/2652524.2652557

[37] K. Petersen and N. bin Ali, "Identifying strategies for study selection in systematic reviews and maps," in *Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement, ESEM*. IEEE Computer Society, 2011, pp. 351–354. [Online]. https://doi.org/10.1109/ESEM.2011.46

[38] C. Wohlin, P. Runeson, P.A. da Mota Silveira Neto, E. Engström, I. do Carmo Machado et al., "On the reliability of mapping studies in software engineering," *J. Syst. Softw.*, Vol. 86, No. 10, 2013, pp. 2594–2610. [Online]. https://doi.org/10.1016/j.jss.2013.04.076