Check for updates

Received: 2 November 2018 | Revised: 7 August 2019 | Accepted: 6 September 2019

DOI: 10.1002/jrsm.1378

RESEARCH ARTICLE

WILEY

# Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources

**Michael Gusenbauer**[1] | **Neal R. Haddaway**[2,3]

[1]Institute of Innovation Management, Johannes Kepler University Linz, Linz, Austria

[2]Stockholm Environment Institute, Linnégatan 87D, Stockholm, Sweden

[3]Africa Centre for Evidence, University of Johannesburg, Johannesburg, South Africa

**Correspondence**
Michael Gusenbauer, Institute of Innovation Management, Johannes Kepler University Linz, Linz, Austria.
Email: michael.gusenbauer@jku.at

Rigorous evidence identification is essential for systematic reviews and meta-analyses (evidence syntheses) because the sample selection of relevant studies determines a review's outcome, validity, and explanatory power. Yet, the search systems allowing access to this evidence provide varying levels of precision, recall, and reproducibility and also demand different levels of effort. To date, it remains unclear which search systems are most appropriate for evidence synthesis and why. Advice on which search engines and bibliographic databases to choose for systematic searches is limited and lacking systematic, empirical performance assessments. This study investigates and compares the systematic search qualities of 28 widely used academic search systems, including Google Scholar, PubMed, and Web of Science. A novel, query-based method tests how well users are able to interact and retrieve records with each system. The study is the first to show the extent to which search systems can effectively and efficiently perform (Boolean) searches with regards to precision, recall, and reproducibility. We found substantial differences in the performance of search systems, meaning that their usability in systematic searches varies. Indeed, only half of the search systems analyzed and only a few Open Access databases can be recommended for evidence syntheses without adding substantial caveats. Particularly, our findings demonstrate why Google Scholar is inappropriate as principal search system. We call for database owners to recognize the requirements of evidence synthesis and for academic journals to reassess quality requirements for systematic reviews. Our findings aim to support researchers in conducting better searches for better evidence synthesis.

**KEYWORDS**

academic search systems, discovery, evaluation, information retrieval, systematic review, systematic search

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2019 The Authors. Research Synthesis Methods published by John Wiley & Sons Ltd

## 1 | INTRODUCTION

Research output, as measured by the number of academic publications, continues to grow exponentially,[1,2] placing scientists in danger of becoming decoupled from the discourse with which they are engaged. The growing volume of research makes it ever harder for practitioners and researchers to keep track of past and current findings in a specific discipline and across disciplines. As a result, research agendas neither build on nor advance previous findings but exist in isolation from the greater body of evidence. Scientific discourse and cumulative knowledge are threatened if researchers fail to connect their empirical or theoretical analyses with past knowledge. As a consequence, the relevance and impact of their research is reduced.[3] In academia, it is true that "we are drowning in information, but starving for knowledge"(4, p12).

Evidence syntheses—the collective term for robust summaries of evidence—aim to mitigate the issues of decoupled empirical evidence amidst ever-growing research output. In particular, high-quality evidence synthesis, in the form of systematic reviews, systematic maps, and meta-analyses (which themselves should be based on systematic searches and critical appraisal), aims "to produce an unbiased description of the cumulative state of evidence on a research problem or hypothesis"(5, p32) and syntheses are thus "viewed as the most reliable sources of evidence for practice and policy."[6] In this way, evidence synthesis is capable of highlighting important developments in a particular field of research.

It is the way in which evidence synthesis is undertaken that "may enhance or undermine the trustworthiness of its conclusion or, in common social science parlance, can create threats to the validity of its conclusions."(5, p33) By following strict rules and guidance, a systematic review provides a comprehensive synthesis of a well-defined area of research. The research team conducting the review must be capable of undertaking online searches using a fit-for-purpose set of search systems, which will enable the researchers to search for and identify *all* available relevant research in a procedurally *unbiased* manner.[7] The constructs and phenomena in question need to be well-defined by the review team so that the online search can use these linguistic cues as frames for searching. Documentation of the search process "makes the search replicable and provides a clear starting point for later updates."[8] Building a systematic review team that incorporates diverse expertise in areas such as content, systematic review methods, searching, and quantitative synthesis has been shown to significantly improve the quality of the review work.[9] Building on an understanding of constructs, a systematic search is the gatekeeper that establishes the basis for subsequent synthesis.[10] This process defines the scope of the examination and thus influences the outcome of the analysis. The same analysis employed with different samples might lead to different results. Similarly, as search systems differ in functionality and characteristics, the same query employed with a different search system may result in a different sample.

Today, evidence synthesis can benefit considerably from innovations in information and communication technology. The introduction of improved tools and methods (also called "evidence-synthesis technology") makes it easier to conduct synthesis work. Technologies such as word processing and reference management software, data analysis, and web-based literature searches allow the more efficient and effective identification, analysis, synthesis, and reporting of research. In particular, online literature search tools now cover most disciplines and have made millions of scientific records searchable within seconds; in some cases, free-of-charge. However, the time saved in physically

**What is already known**

- Evidence identification in systematic reviews and meta-analyses requires the right search strategy using the right search systems.
- Until now, researchers have lacked comprehensive guidance on which search systems are suitable for systematic searches.

**What is new?**

- This study provides a systematic evaluation and comparison of search and retrieval qualities of 28 widely used academic search systems, including Google Scholar, PubMed, and Web of Science.
- Evaluation profiles for each of those 28 systems allow researchers to assess why and to what degree a particular system is suitable for their search requirements.

### Potential impact for RSM readers outside the authors' field

- By making qualities and limitations of search systems transparent, this study creates awareness across disciplines among journals and among database providers to pay particular attention to the search requirements of evidence synthesis.
- We hope our findings assist researchers to perform better searches that require less time, identify more relevant evidence, and adhere to systematic review guidance.

searching for evidence must now be spent in carefully planning searches that make use of complex syntax and search facilities to mine the textual data within titles, abstracts, keywords, and full texts. Accordingly, collaboration with librarians, as information experts in these complex literature search processes, has been shown to benefit study quality.[9,11,12] In evidence syntheses, online databases should "form the backbone of any comprehensive literature search. These sources probably contain the information most closely approximating all research."(5, p112) Indeed, it is now impossible to imagine undertaking academic work without using web-based literature search systems.

Rigorous evidence syntheses, such as systematic reviews, have specific requirements for literature searches.[13] These requirements are stipulated in *conduct* guidance issued by renowned institutions dedicated to warrant and elevate the quality of evidence synthesis in academia. We have based our further analysis on guidance published by three institutions: Cochrane, The Campbell Collaboration, and the Collaboration for Environmental Evidence (CEE). We decided not to include PRISMA and ROSES guidance, as these resources offer guidance on *reporting* rather than *conduct*. Below, we provide an overview of how searches for studies to be included in the evidence base should be performed in systematic reviews.

**TABLE 1** Quality requirements of systematic searches derived from evidence synthesis guidelines

| Source | Quality Requirements |
|---|---|
| Cochrane Handbook, 2011 | "The key characteristics of a systematic review are: [...] an explicit, reproducible methodology; a systematic search that attempts to identify all studies that would meet the eligibility criteria [...]."[14] |
| Campbell Methods Guides, 2016 | "Systematic reviews of interventions require a thorough, objective and reproducible search of a range of sources to identify as many relevant studies as possible (within resource limits)."[15] |
| CEE Guidelines and Standards for Environmental Evidence Synthesis, 2018 | "To achieve a rigorous evidence synthesis searches should be transparent and reproducible and minimise biases. A key requirement of a review team engaged in evidence synthesis is to try to gather a maximum of the available relevant documented bibliographic evidence in articles and the studies reported therein."[16] |

Guidance for systematic reviews (Table 1) refer to three recurring quality requirements that are critical for literature searches: First, the goal must be to identify *all relevant records* (or as many as the resources of the reviewer permit). Second, the search must be *transparent*. Third, the search must be *reproducible*.

Reviewers must take great care when following the steps of the systematic review in order to meet quality requirements. The choice of one or more adequate search systems that allow the user to meet these quality requirements is one major consideration because these systems determine the number of relevant articles that will be identified. If a system has limitations of some sort, even the most skillful searcher might find them difficult or impossible to circumvent. Because the search systems differ in technical characteristics and scope, their suitability for systematic searches and thus for systematic reviews varies; however, reviewers are often unaware of the technical characteristics and limitations of search systems. Reviewers consulting methods-guidance on search systems will find only advice pertaining to certain systems; advice which is often not based on a systematic review of search functionalities. Accordingly, review efforts can still benefit significantly from guidance on which search systems are most suitable for a specific search task.

The first goal of a systematic review is to identify all or as many as possible relevant resources. Hence, the reviewer needs to select a search system that provides the best *coverage* of the chosen search topic. Coverage of a search system is denoted relative to a specific criterion—for example, a specific subject (eg, medicine and physics), resource type (eg, articles and books), time span (eg, retrospective coverage), or geographic location. A search system might provide high coverage of articles in medicine, yet low coverage of articles in physics. Greater overall size of a search system in this sense does not necessarily denote greater coverage on a specific topic. For example, while the multidisciplinary search system JSTOR has more than 12 million records and is considerably larger than IEEE Xplore with four million records, the coverage of IEEE Xplore on the specific topic of engineering records is still broader and thus more appropriate for evidence synthesis in engineering. Accordingly, systematic review guidance advises the use of suitable specialized databases that provide high coverage of a specific topic as well as generic resources that have broad coverage. Reviewers should thus consider their specific review topic when deciding which search systems might prove suitable for a systematic search. To assist with selection, there is considerable research on the coverage of search systems,[17-19] especially with regard to search systems such as Google Scholar which have built up an aura of secrecy around the size of their databases.[20-22]

While coverage is an important criterion based on the specific requirements of the systematic review, it does not indicate how well a reviewer can in fact access these resources on the chosen search system. A search system must allow a reviewer to specify queries that search with high *recall* and *precision*. While recall (or sensitivity) is the percentage of relevant article records that are returned in the result set from all relevant records known to exist, precision (or specificity) is the percentage of records in the result set that are relevant.[23,24] While high recall indicates a search contains many of the relevant items, high precision indicates a search retrieves relatively few irrelevant records.[5] Generally, the more recall is improved, the greater the reduction in precision, and the more precision is improved, the worse the effect on recall.[23] "Although precision and recall are typically at odds, there's one way to overcome the constraints of this trade-off: more features."(23, p80) More features in this context means that recall and precision can both be improved if the search query of a reviewer can be refined so it more accurately includes relevant records, while excluding irrelevant records. The ability of a reviewer to manipulate their search query depends on the capabilities of a search system and thus they significantly influence recall and precision.[25] The capability of search systems to retrieve results in an effective and efficient manner determines its suitability in systematic searches. A suitable search system provides good coverage in the specific area of interest and allows the user to specify a query with high precision and recall. Accordingly, search systems "should be evaluated against the background of what is found for—and what remains hidden from—the users."(26, p1570)

While the goal is to identify all records on a given topic, in practice, this goal has to be pursued within *resource limits*. Reviewers thus must make reasonable judgements on where to best invest time and funds based on a cost/benefit analysis: "Searches for systematic reviews aim to be as extensive as possible in order to ensure that as many as possible of the necessary and relevant studies are included in the review. It is, however, necessary to strike a balance between striving for comprehensiveness and maintaining relevance when developing a search strategy. [...] The decision as to how much to invest in the search process depends on the question a review addresses and the resources that are available."(15, p26) As a consequence, reviewers must select search systems that allow them to make the best use of their resources, that is, to retrieve the most relevant records in exchange for the least amount of time or funds.

The second and third goal of systematic reviews, *reproducibility* (also "replicability," "reliability," and "repeatability") and *transparency*, require an explicit, transparent, and documented search process that allows reviewers to update or replicate a given synthesis search. "The search process needs to be documented in enough detail throughout the process to ensure that it can be reported correctly in the review, to the extent that all the searches of all the databases are reproducible."(15, p41) Conduct and reporting guidance explicitly describe the steps necessary for a reviewer to ensure the rigorous and transparent documentation necessary to foster reproducibility. However, the functionality and capabilities of specific search systems can also influence reproducibility themselves. The reproducibility of a search determines whether that search can be replicated employing the same methods with a given search system. If the same query leads to the same search results, the search is considered reproducible. A lack of reproducibility can indicate a form of sampling bias or so-called search engine bias[27,28]: Repeated tests of the same query can lead to different results.(29, p37) It is important to note that because the size of the database provided by a search system typically increases over time, repeated queries will naturally yield a larger set of results than the initial query, and this has to be taken into account in assessing search system reproducibility and should not be seen as a lack of reproducibility. Researchers must thus pay close attention to reporting the date on which searches took place to make changes in the database of the search system comprehensible.

In summary, it is important to consider how far a search system supports the user in articulating and framing a query in a systematic search context, with special attention to high levels of coverage, recall, precision, and reproducibility.

The objective of the current research is to test and describe the usability and functionality of search systems that are frequently used in evidence syntheses, focusing on limitations that may influence the quality of systematic reviews. While some of these limitations impede the rigor required for systematic reviews as laid out by methodological guidance (necessary condition), others make systematic reviews more challenging or resource intensive (desired condition). Failing at some necessary condition does not mean the search system should be avoided entirely in the systematic review process, but does mean it should perhaps not be used for query-based searching: the fundamental underlying search method for systematic reviews. Nevertheless, such systems might be used in supplementary search methods.

Previous studies examining the suitability of search systems for evidence synthesis have focused on a limited number of search systems and/or based their analysis on a review of the search interface, yet without any in-depth examination of core functionalities that allow reliable query-based searching.[19,30-32] Previous studies have also calculated precision and recall of search systems from data reported by specific evidence-synthesis

**TABLE 2** Description of tests performed to evaluate academic search systems, based on systematic search requirements regarding coverage, recall, precision, efficiency, and reproducibility

| Number | Tested Quality | Test Scope | Test Criterion | Test Procedure | Performance Threshold of Test Criterion | Necessity of Meeting Threshold |
|---|---|---|---|---|---|---|
| 1 | C, R, P, E | DB | Subject coverage | Review of content description: type of academic disciplines covered | Subject coverage: maximum | Desired |
| 2 | C, R, P, E | DB | Size | Review of information provided by the official website of each search system concerning number of records provided on a database.<br>Estimations of sizes based on query method by Gusenbauer, 2019[37]. | Size: maximum | Desired |
| 3 | C, R, P, E | DB | Record type | Review of types of scholarly resources offered. | Types of records: maximum | Desired |
| 4 | C, R, P, E | DB | Retrospective coverage | Limitation of time span to oldest years available. | Time span: maximum | Desired |
| 5 | C, R, P, E | DB | Open access content | Review of search system and description of content provided: type of usage rights attributed to resources. | Content: open | Desired |
| 6 | R, P, E | Q | Controlled vocabulary? | Review of search options. Is a controlled vocabulary available and what is its coverage and accessibility?<br>1. Available: yes/no<br>2. Retrospective coverage: years available<br>3. Hierarchically structured: yes/no<br>4. Searchable: yes/no | Available: yes | Desired |
| 7 | R, P, E | Q | Field code search: query refinement | Listing of all field codes and limiters visible in the search interface. | ≥5 field codes | Necessary |
| 8 | R, P, E | Q | Full text search option available? | Review of search options (field codes). | Available: yes | Desired |
| 9 | R, P, E | Q | Search string:<br>1. maximum length of word combination<br>2. maximum number of characters | 1. maximum search string length (trial and error)<br>2. maximum number of characters (web search) | ≥25 terms | Necessary |
| 10 | R, P, E | Q | Server response (time and number of records) for max. word combination | Test of the longest search string the system still could handle and review of search results in terms of whether longer search strings produced more results.<br>1. test via maximum word combination from test 9<br>2. test of next shorter word combination | Longer search string = more hits<br>Timeout: no | Necessary |

(Continues)

**TABLE 2** (Continued)

| Number | Tested Quality | Test Scope | Test Criterion | Test Procedure | Performance Threshold of Test Criterion | Necessity of Meeting Threshold |
|---|---|---|---|---|---|---|
| 11 | R, P, E | Q | Search string language support: English, Chinese, Cyrillic | 1. English tested with test number 9<br>2. Chinese tested with 的 (means "bright")<br>3. Cyrillic tested with и (single letter) | Support: full (Chinese and Cyrillic) | Desired |
| 12 | R, P, E | Q | Boolean "OR" functional? | Test of six terms (research, define, paper, Asterix, table, and analysis) through step-wise addition with "OR" operators → does record count increase? | Functional: yes | Necessary |
| 13 | R, P, E | Q | Boolean "AND" functional? | Test of six terms (research, define, paper, Asterix, table, and analysis) through step-wise addition with "AND" operators → does record count decrease? | Functional: yes | Necessary |
| 14 | R, P, E | Q | Boolean "NOT" functional? | Test of six terms (research, define, paper, Asterix, table, and analysis) through step-wise addition with "NOT" operators → does record count decrease? | Functional: yes | Necessary |
| 15 | R, P, E | Q | Comparative test: "AND"/"NOT"comparative test: "OR"/"NOT" | *Comparative test* of AND/NOT-operators: Does "research" minus "research AND define" equal "research NOT define"?<br>*Comparative test* of OR/NOT-operators: Does "define" plus "research NOT define" equal "research OR define"? | Functional: yes | Necessary |
| 16 | R, P, E | Q | Literal vs. expanded queries | Comparison of various similar (ill-written) terms:<br>1. Defin, "Defin"<br>2. Definx, "Definx"<br>3. Define, "Define"<br>Do databases interpret queries literally, ie, retrieve only few hits for ill-written terms? Do quotation marks (usually used as limiters to search for verbatim) reduce the number of results, indicating that queries without quotation marks were automatically expanded?<br>Comparison of British and American English:<br>4. Organise, "Organize"<br>5. Colour, "Color"<br>Does variation in British/American spelling lead to differences in search results. If not, queries are expanded automatically to include both British and American versions. | Literal queries: yes | Desired |

(Continues)

**TABLE 2** (Continued)

| Number | Tested Quality | Test Scope | Test Criterion | Test Procedure | Performance Threshold of Test Criterion | Necessity of Meeting Threshold |
|---|---|---|---|---|---|---|
| 17 | R, P, E | Q | Truncation/wildcards available? | Use of most common symbols:<br>1. "*" for right-hand truncation<br>2. "?" for single character placeholder<br>3. "$" for zero- or one-character placeholder | Functional: yes | Desired |
| 18 | R, P, E | Q | Exact phrase search functional? | Exact phrase search was tested using "":<br>1. Organise team, "Organise team"<br>Does the use of quotation marks change search results? | Functional: yes | Necessary |
| 19 | R, P, E | Q | Parenthesis functional? | Scope tested with different positions of parentheses:<br>1. (Research OR define) AND Asterix<br>2. Research OR (define AND Asterix) | Functional: yes | Necessary |
| 20 | R, P, E | F | Filtering: Post-query refinement | Listing of all filters that allow refinement of a query results set visible in the search interface. | Number of filters: maximum | Desired |
| 21 | R, P, E | CS | Forward citation search available? | Review of search options. | Available: yes | Desired |
| 22 | E | Q | Advanced search string input field | Review of search options. | Available: yes | Desired |
| 23 | E | Q, F, CS | Search help? | Review of search options. | Available: yes | Desired |
| 24 | R, E | Q, F, CS | Maximum number of accessible hits | Last accessible results page of a large results set. | ≥1000 results | Necessary |
| 25 | E | Q, F, CS | Bulk download supported? | For how many records can citation information be downloaded at once (involving a single selection and download request) to be exported to some reference management software or other destination? | Supported: maximum | Desired |
| 26 | RP | Q | Reproducibility of search results at different times | Repeated query after few seconds. | Search result: same | Necessary |
| 27 | RP | Q | Reproducibility of search results at different locations | Repeated query after few seconds with different, foreign IP address, or different institutional access. | Search result: same | Necessary |

Abbreviations: C, coverage; CS, citation search; DB, database; E, efficiency; F, filter; P, precision; Q, query; R, recall; RP, reproducibility.

studies.[24,32-34] This analysis focuses instead on evaluation criteria for systematic reviews across disciplines, following universally accepted conduct guidance (ie, Cochrane, Campbell, and CEE). This research thus fills a need for support in the choice of search system, currently lacking in evidence-synthesis methodology, and follows calls for comparative studies on the effectiveness of search systems[25] or the "need to develop 'bias profiles' for search engines."(35, p1193) This study provides a much-needed overview of academic search systems from a user perspective. It compares a large selection of popular academic search systems and examines their unique characteristics to draw conclusions on their suitability for evidence synthesis. Specifically, the study tests whether a system allows the user to precisely specify a search so it retrieves as many relevant results as possible, how efficiently search results can be retrieved, and if the search results could be reproduced with the same methods. Hence, our study contributes to evidence synthesis as "[...], the value of a systematic review depends on what was done, what was found, and the clarity of reporting."(36, p1) Overall, the question framing this research is: How suitable and usable are commonly-used academic search systems for systematic searches in evidence synthesis? Definitions of the terms used throughout this study can be found in Appendix I, the detailed tests in Appendix II (supplementary online material).

## 2 | METHOD AND ANALYSIS

This study measures the suitability of a number of popular search systems for evidence synthesis using specific criteria. These criteria were assessed based on the 27 tests outlined in Table 2.

### 2.1 | Selection of search systems

The search systems analyzed in this study represent common resources in highly cited systematic reviews and meta-analyses in recent years. According to a search of Web of Science, for systematic reviews and meta-analyses across all databases available to us,* there were 63 "hot papers" that were "published in the past two years and received enough citations in September/October 2018 to place it in the top 0.1% of papers in its academic field." All search systems and databases that were mentioned in at least two of these 63 studies were included in our analysis. The result was a list of 16 databases and search systems: CINAHL, ClinicalTrials.gov, Cochrane Library, EbscoHost, Embase, ERIC, Google Scholar, LILACS, ProQuest, PsycINFO, PubMed, ScienceDirect, Scopus, SportDiscus, TRID, and Web of Science. While most of the search systems mentioned were databases, some authors mentioned platforms without stating the exact databases searched (eg, Web of Science is a platform, while Web of Science Core Collections is its main database)—a common reporting error of search scope.

In addition, to obtain a broader picture of the qualities of academic search systems, we also included other search systems that are regularly used among academic researchers across disciplines[38]: AMiner, ACM, arXiv, Bielefeld Academic Search Engine (BASE), CiteSeerX, Digital Bibliography & Library Project (DBLP), Directory of Open Access Journals (DOAJ), IEEE Xplore Digital Library, JSTOR, Microsoft Academic, Semantic Scholar, SpringerLink, Wiley Online Library, WorldCat, and WorldWideScience. Thus, we examine the quality of a total of 28 search systems that access 34 databases either via web search engines (eg, Google Scholar or Microsoft Academic), via platforms that allow access to one or more discrete databases (eg, ProQuest or OVID) or other bibliographic databases (eg, Transport Research International Documentation). Below, we present an overview of the 28 search systems; if the database is accessed via a platform, the database's name is given in parentheses as follows:

| | | |
|---|---|---|
| 1. ACM Digital Library | 11. Education Resources Information Center | 21. Semantic Scholar |
| 2. AMiner | 12. Google Scholar | 22. SpringerLink |
| 3. arXiv | 13. IEEE Xplore Digital Library | 23. Transport Research International Documentation |
| 4. Bielefeld Academic Search Engine | 14. JSTOR | 24. Virtual Health Library *(LILACS)* |
| 5. CiteSeerX | 15. Microsoft Academic | 25. Web of Science *(Medline, Web of Science Core Collection)* |
| 6. ClinicalTrials.gov | 16. OVID *(Embase/Embase Classic, PsycINFO)* | 26. Wiley Online Library |

(Continues)

*Web of Science Core Collection, BIOSIS Citation Index, BIOSIS Previews, Data Citation Index, Derwent Innovations Index, KCI-Korean Journal Database, MEDLINE, Russian Science Citation Index, SciELO Citation Index, and Zoological Record (accessed on February 11, 2019; search string: ti("systematic review" OR "meta-analysis")

| | | |
|---|---|---|
| 7. Cochrane Library *(CENTRAL)* | 17. ProQuest *(ABI/Inform Global, Nursing & Allied Health Database, Public Health Database)* | 27. WorldCat–Thesis/Dissertations |
| 8. Digital Bibliography & Library Project | 18. PubMed *(Medline)* | 28. World WideScience |
| 9. Directory of Open Access Journals | 19. ScienceDirect | |
| 10. EbscoHost *(CINAHL Plus, EconLit, ERIC, Medline, SportDiscus)* | 20. Scopus | |

We selected a large set of popular specialized and multidisciplinary search systems that are relevant not only for disciplines where evidence synthesis is already well-established (eg, medicine, health sciences, or environmental studies) but also other disciplines, such as management, where these methods have been increasingly used just in the last years. To include all search systems or databases in this study would be an impossible task, as hundreds of bibliographic databases and search systems exist across subjects.

Our sample of search systems covers a range of types of technology (eg, platforms and web search engines), target audience (eg, academic discipline and resource restriction), and provided content (eg, traditional academic literature and grey literature). Some platforms examined provide access to multiple databases at once, allowing us to assess the basic qualities and functionalities on a system level as well as a database level. While we find that most of the functionalities are determined by the system itself, other qualities might be closely linked to the underlying database—such as the number and type of field codes or the availability of a controlled vocabulary. We included proprietary, nonproprietary, and Open Access databases, a distinction especially relevant for reviewers who have only limited access to expensive database subscriptions. While the focus was on bibliographic databases, we also included sources that include grey literature (eg, arXiv, Google Scholar, and WorldCat-Thesis/Dissertations). Grey literature refers to any document produced by an organization at any level whose primary purpose is not commercial publishing and includes theses, white papers, organizational reports, and consultancy documents.[39] By searching for grey literature, systematic reviews aim to maximize comprehensiveness and mitigate publication bias.[16] Typically, systematic reviews will conduct dedicated searches for grey literature (for example, searching organizational websites), but the ability to include grey literature in formalised, systematic searches of bibliographic databases can provide benefits, including the ability to assess eligibility concurrently with bibliographic search results, potentially increasing efficiency.

## 2.2 | Evaluation approach: Different search systems—One overarching method

It is important to note that the search systems we analyze in our sample are diverse in their functionality, syntax, and features. All of these systems have different underlying databases and indexing methods, data presentation, and curation methods. Crawler-based web search engines (eg, Google Scholar), for example, function differently from bibliographic databases which have a curated catalogue of information (eg, Scopus). Some of these search systems are large and multidisciplinary (eg, Scopus), while others have a narrower focus on a single or a few domains of research (eg, PsycINFO) (see Appendix II). We examine these diverse search systems through the lens of the users that access them and test how well the search facility performs to link the query to the underlying database. We do not test the searchers' ability to formulate such strings/queries[40-44] and we do not test the completeness of the underlying dataset provided by the search system.[24,45] This study instead examines the search system as the gatekeeper that mediates between a database of potentially relevant records and a reviewer that wishes to access, retrieve, analyze, and synthesize that information in a systematic, rigorous manner.

Hence, we assessed the functionality of these databases with standard queries from the perspective of the user (see Table 2). In querying diverse databases with a diverse set of inputs, we tested the capacity of the databases to interpret the user's query so that the dataset is retrieved effectively. We examined the results of the search systems both quantitatively (eg, how many hits a query retrieved or how much time the server needed to respond) and qualitatively (eg, the nature of the search options and the search interface). The quantitative methods based on tested methods that use variations of search queries to iteratively determine sizes of different types of search systems.[37] In our analysis, we did not assess the quality of the retrieved records, in terms of their fit with a given search intent for example. Quality

**TABLE 3** Review of search methods used in systematic reviews

| Number | Study | Search System with Largest Result Set | Use of Field Codes | Length of Longest Search String (Boolean Operators, Field Codes Not Counted) | Use of Boolean Operators (AND, OR, NOT) | Use of "" | Use of () | Use of Truncation, Wildcards | Total Number of Studies Screened (Incl. Duplicates) | Studies Identified with a Single Search String | Non-query Identification (Total) | Size of Final Sample | Search Precision |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Aune et al[58] | PubMed, Embase | N/A | 68 terms | OR, AND | Yes | Yes | No | 49 772 | 40 744 | 4 (handsearch, all relevant) | 142 | 0.28% |
| 2 | Barnett et al. (2017)[59] | Scopus | N/A | N/A | N/A | N/A | N/A | N/A | 19 005 | 5681 | 3 (handsearch, all relevant) | 100 | 0.53% |
| 3 | Baur et al[60] | PubMed | Yes | 57 terms | OR, AND | Yes | Yes | No | 1169 | 1113 (sum of all databases) | 56 (handsearch, not all relevant) | 76 | 6.50% |
| 4 | Bediou et al[61] | N/A | N/A | 11 terms | OR, AND | Yes | Yes | No | 958 147 | N/A | 0 | 82 | 0.01% |
| 5 | Bethel et al[62] | PubMed | N/A | 16 terms | OR, AND | Yes | N/A | No | 12 | N/A | N/A (post-query filtering) | 4 | 33.33% |
| 6 | Bourne et al. (2017)[63] | Embase (via OVID) | Yes | 12 terms (sets were later combined) | Or, And,not | Yes | Yes | Yes | 3878 | 2539 | N/A (post-query filtering) | 288 | 7.42% |
| 7 | Brunoni et al[64] | PsycInfo (via OVID) | Yes | 10 terms | OR, AND | Yes | Yes | Yes | 1121 | N/A | N/A | 81 | 7.23% |
| 8 | Carlbring et al[65] | PubMed | N/A | 29 terms | OR, AND | Yes | Yes | No | 2078 | 2078 | N/A (handsearch, post-query filtering) | 20 | 0.96% |
| 9 | Chu et al[66] | Medline, Healthstar (via OVID) | Yes | 25 terms (sets were later combined) | OR, AND | Yes | Yes | Yes | 1784 | N/A | N/A | 26 | 1.46% |
| 10 | Cipriani et al[67] | N/A | N/A | 10+ terms (combined with undisclosed keyword list) | OR, AND | Yes | Yes | Yes | 28 552 | 24 200 (sum of all databases) | 4352 (handsearch, other sources) | 421 | 1.47% |
| | Recommended threshold? | | | ≥25 terms | OR, AND, NOT | Yes | Yes | Yes | | | Post-query | | |

(Continues)

**TABLE 3** (Continued)

| Number | Study | Search System with Largest Result Set | Use of Field Codes | Length of Longest Search String (Boolean Operators, Field Codes Not Counted) | Use of Boolean Operators (AND, OR, NOT) | Use of "" | Use of () | Use of Truncation, Wildcards | Total Number of Studies Screened (Incl. Duplicates) | Studies Identified with a Single Search String | Non-query Identification (Total) | Size of Final Sample | Search Precision |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ≥5 field codes | | | | | | | ≥1000 accessible records | filtering: m aximum Citation search: yes | | |
| | Necessity? | | N | N | N | N | N | D | | N | D | | |

Abbreviations: N, necessary; D, desired.

criteria have previously been used to evaluate search systems in terms of recall and precision indicating their suitability of search in general and systematic search in particular.[25,33,46-57] If available, we searched with the advanced search interface of the search system. Tests were performed between February and March 2019.

## 2.3 | Necessity to meet requirements

Our evaluation of search systems involves applying 27 unique criteria each of which tests the performance of a specific quality. In our evaluation of the search systems, we differentiate between capabilities that are *necessary* or merely *desired* for a systematic review. In order to meet the requirements of the guidance of Cochrane, The Campbell Collaboration, and CEE for systematic reviews, a necessary criterion needs to be fulfilled by a search system, irrespective of the context of the study. A desired criterion is necessary to be met only for systematic reviews with specific requirements or foci. Further, a desired criterion that is not met, can, if the reviewer is aware, be circumvented with extra effort of using suboptimal search methods. Reviewers should decide whether the fulfilment of a desired criterion is important for their specific search. Necessary criteria, however, should always be met by search systems. Each criterion was classified as either desired or necessary according to evidence synthesis guidance (see Table 2).

In order to support our decision to determine meaningful performance thresholds, we reviewed the search methods of 10 random articles from the sample of 63 - articles (see Table 3). We reasoned search systems should at least come close to enabling the searches described in these studies. We extracted information concerning the search methods these studies used to obtain their search results. The single thresholds are explained in detail in the description of each single criterion used in our test.

## 2.4 | Requirements translated to evaluation criteria of search systems

The requirements for systematic reviews are largely agreed upon in evidence-synthesis guidance: (a) identify a maximum number of relevant records for a specific topic, (b) within the resource limits, and (c) use transparent and reproducible search methods. These requirements focus on the search process of the reviewer and the type of methods that are employed; yet, evidence-synthesis guidance provides no clear technical requirements for search systems. Hence, for our analysis, we

translated these requirements to technical criteria that search systems needed in order to meet the requirements of evidence synthesis. A search system thus needs to be (a) *effective* in finding most of the relevant results while filtering out the irrelevant, (b) *efficient* allowing the reviewer fast identification and retrieval of records, and (c) must allow the *reproduction* of search results with the same methods:

First, effective search depends in the reviewer's choice of a suitable search system offering the best coverage of records searched for. Coverage can be determined in multiple ways: for example, concerning time frame (retrospective coverage), academic subject (discipline) or usage rights (Open Access). Additionally, *effectiveness* is determined by the search system's capability to translate the search frame determined by the reviewer to enable precise searching with a high level of recall. To provide a thorough search for relevant records, the reviewer can combine different methods of (a) queries with keywords and a controlled vocabulary, (b) post-query filtering, and (c) handsearching of relevant journals, issues, or reference lists (citation search). While all of these methods provide value for a rigorous search that aims to identify all or at least most relevant records, it is particularly *queries* using keywords or a controlled vocabulary that are able to search the corners of a database that would be inaccessible with citation searching, handsearching, or post-query filtering alone. The Cochrane Review of Stacey et al[68] is an example of a study that uses an elaborate query-based search strategy relying on "AND," "NOT," "OR" operators, database-specific field codes and controlled vocabularies. Searching five databases from different providers, they retrieved a total of 46 054 hits from which they included 105 studies in their final meta-analysis. A perfect query would, in theory, make citation searching and handsearching obsolete, yet perfect citation searching and handsearching could not do the same to make the use of queries obsolete. This logic is supported by our review (see Table 3) where most identified results were derived from query-based searches—some studies based their search strategies on queries alone. Accordingly, using *queries* for systematic search is necessary for systematic reviews as evidenced in both research practice and in evidence-synthesis guidance. The other search methods are supplementary techniques desired to improve the search result of the query. Second, the efficiency with which a reviewer can retrieve relevant results is largely determined by recall and precision. Therefore, the choice of a suitable search system with suitable coverage and capabilities of searching with functional search strings, filters, and citation search impacts tremendously on effectiveness, that is, precision and recall. Nevertheless, other functionalities associated with downloading search results or user-friendly data-input also influence search efficiency (along with the subsequent stage in a systematic review, eligibility assessment). Third, *reproducibility* can be determined how well the system is capable to retrieve same results again with same search methods.

## 2.5 | Test procedures and performance requirements

We reviewed and tested the 28 search systems with 27 criteria determining each search system's (a) coverage and (b) capability to perform systematic searches via queries, filters, and handsearching so that a reviewer can obtain *reproducible* results, *efficiently*, and with high recall and precision.

### 2.5.1 | Coverage

Generally, it is assumed that more coverage is better than less coverage, as without a comprehensive database, searches would identify few relevant records. This means higher coverage typically increases the recall of a query. Nevertheless, it is important to note that while recall increases, precision simultaneously decreases, disproving the statement that more coverage is *always* better. What records are considered relevant depends on the specific requirements of the reviewer and thus cannot be generalised. For example, a search system with a smaller size, covering only a single discipline, might bring more relevant search results than a large search system covering multiple disciplines. Accordingly, all performance requirements on coverage were framed as *desired* criteria as reviewers must decide for themselves what search system best fits their unique study requirements.

*Criterion 1: "Subject coverage"* assesses the type of academic disciplines that are predominantly covered by a search system. This criterion determines whether a search system specializes in single disciplines or is multidisciplinary. While a greater coverage of disciplines might generally be regarded as beneficial, the greater breadth of records available might harm search precision. When working with such multidisciplinary search systems, the reviewer needs to be more specific about search context to receive the same precision than when using a specialized search system.

*Criterion 2: "Size"* informs about the absolute number of records available on a database that is made available through a search system. Searching larger databases, all things being equal, results in higher search recall. We assessed sizes by reviewing the official information

provided by the search systems' websites. If this information was up-to-date, we reported the official number; if it was either outdated or unavailable, we used the method suggested by Gusenbauer[37] to assess search system size.

*Criterion 3: "Record type"* informs about the types of records offered by a search system. Here we relied on the information provided by the search systems. Naturally, each search system had its own definition of how to categorize and classify records, which made direct comparison of record types difficult. Nevertheless, the availability of more—as opposed to fewer—document types meant that reviewers could search with increased precision, if this field code was available to specify a search.

*Criterion 4: "Retrospective coverage"* informs about what year the oldest records on a database are from. When information on retrospective coverage was provided by the search system, we included this information, if not, we manually searched for the oldest records on the database and reported this year. In doing so, we took care not to consider incorrectly dated records in our assessment of retrospective coverage.

*Criterion 5: "Open Access"* was assessed in reviewing the usage rights of the records offered. If a search system offered mostly proprietary content with only a marginal focus on Open Access resources, then it was considered "proprietary." If the search system was nonprofit and/or provided strong emphasis to support Open Access content—but was also linked to proprietary resources—it was considered "mixed." If the search system offered only Open Access content, it was considered "open."

### 2.5.2 | Search

All of our query tests (criteria 6-17) have in common that the results determine whether a given search system allows the user to specify a systematic search query that is targeted at high precision or recall. All these query-features are helpful to compile a comprehensive search string to specify exactly what lies inside and outside the search scope of the evidence synthesis.

*Criterion 6: "Controlled Vocabulary"* was assessed by reviewing the search options provided by the search system for a given database. Where databases are accessed via broader platforms (eg, ProQuest), the options available often differed across databases, such that a single platform may have different search functionalities for its databases. The controlled vocabulary is more useful for some disciplines, while it is less useful for others. For example, "databases in the social sciences tend not to be as thoroughly indexed as those in medicine and may use methodological indexing inconsistently, if at all."(15, p33) Hence, guidance by Campbell Collaboration, for example, advises for cautious use of controlled vocabularies: *"When searching for studies for a systematic review, [...] the extent to which subject terms are applied to references should be viewed with caution. Authors may not describe their methods or objectives well and indexers are not always experts in the subject areas or methodological aspects of the articles that they are indexing."*(15, p28) Further, because retrospective coverage of the controlled vocabulary may be limited, reviewers should take care if relying on this method, especially for searches over longer time periods including earlier studies. It is difficult to quantify the quality of such controlled vocabulary as their features are diverse. Accordingly, we provide additional information on coverage, the option of a searchable index, and the availability of a hierarchical structure. We leave it to reviewers to decide whether such information is helpful for their specific search tasks. In summary, we regard the availability of a controlled vocabulary a desired condition as a thorough query-based search can compensate for some of the advantages of a controlled vocabulary.

*Criterion 7: "Field Code Search: Query Refinement"* is important for systematic searches to provide the user with options to search with high precision and recall. We reviewed the search options to assess which field codes were provided by search systems allowing reviewers to detail which parts of documents should be searched. The availability of field codes is a necessary criterion, since the user must be able to specify exactly where the requested information is located within the records. We defined *five* field codes as the lower threshold. With the exception of the option of a full text search (criterion 8), we do not test for the availability of single field codes, as we assume that in general, the more field codes are available the better the chances of a reviewer being able to search with high recall and precision.

*Criterion 8: "Full Text Search"*. In some cases, reviewers need to search the full texts to identify specific study types. As full text search is however not necessary in every systematic search, we considered this criterion desirable.

*Criterion 9: "Maximum Search String Length"* was an essential determinant of how long and thus how specific the search string can be. We determined the maximum length of search strings that still retrieved results but did not result in timeouts or other system failures through trial and error of search strings varying in length. For this purpose, we used the 2008 Oxford Word List[69] and interlinked strings of the top 1000, top 500, top 100, top 50, top 25, top 10, or fewer most utilized English words with Boolean "OR"-operators. It happened that some Boolean queries were aborted due to timeouts or search string length limitations. Here, we iteratively searched

for the largest query the search system could still handle. We considered a Boolean search needs to support at least *25 terms*, so the user is able to specify the minimum necessary outlines of a search, i.e. to adapt the search string to the linguistic particularities of the scientific (sub-) topics of interest using keywords. Our quick review of systematic searches (Table 3) revealed that while appropriate search strings can be significantly longer, 25 terms should allow reviewers to specify their search scope to a reasonable extent. Search string length is critical for searching with high recall and precision and thus a necessary criterion. If reviewers needed a larger search string than supported by the search system, it might be possible to circumvent this limitation by splitting search strings. This practice can, however, be extremely laborious and is prone to error.

*Criterion 10: "Server Response (Time and Number of Records)"* was a test to determine the server's response time for the longest search string still supported by the system and was conducted together with the test for criterion 9. Additionally, we tried the next shorter search string combination to determine whether longer search strings actually resulted in longer loading times. Further, in order to pass this test, the system needed to produce more results for broader searches than for narrower ones. As the technical performance of the search system and the correct interpretation of long search strings are critical for information retrieval, we deemed this test necessary.

*Criterion 11: "Search String Language Support"* tested the search systems' capacity to interpret English, Chinese, and Cyrillic characters using frequently used terms and characters to determine the response of the search system. If the system retrieved results, it was deemed to support such characters, if the result was an error message or zero results, we deemed the system was not working. This was considered a desired criterion, as reviewers typically search with English characters only and because a number of databases index non-English records using translated English language titles and abstracts, thus being identifiable also with English language.

*Criteria 12 to 15: "Boolean Functionality"* were some of the most important tests in our study and tested the search systems' capability to effectively interpret the most common Boolean operators OR, AND, and NOT.[55] The system must retrieve results as anticipated by the Boolean logic. Boolean operators (AND, OR, NOT) are an integral part of systematic searches,[26] allowing the user to precisely specify the scope of the query as no other technique could. While AND and OR are used to link single concepts to a common search string, NOT is mostly used for disambiguation. Because evidence synthesis has very specific information needs determined by choices of constructs, methods, and research questions, it needs complex search strings to determine queries that link concepts of interest. Boolean operators have been shown to be essential for sampling in evidence synthesis[70] as they "allow the searcher to use set theory to help define the items that will be retrieved by a search."(5, p103) They provide "a great range of strategies are available to increase 'recall' and 'precision.'"(26, p1570) We used a combination of a total of six terms—research, define, paper, Asterix, table, and analysis—to see whether the result set increased, decreased, or remained constant after adding one more term. For strings with OR operators the results set should increase or at least remain constant with each additional term, with AND operators, it should decrease or remain constant, and with NOT, it should decrease or remain constant. We deliberately chose words from the research context and one—the word "Asterix"—that is very unlikely to appear frequently in scholarly articles to test the systems' responses. Adding the nonscholarly term Asterix to a Boolean OR string would not add many additional hits on a scholarly database yet would reduce the set of an AND string to almost zero or decrease it only slightly in a NOT combination. We also tested how many results the term Asterix would retrieve when searched as a single term to cross-check these results against the changes in the different Boolean strings. Further, we used alternative notations for Boolean operators if they were explicitly stated in the help or FAQ files of the individual search systems. We confirmed that "how websites are represented and the precise commands used to do the Boolean syntax search will differ somewhat for each search engine."(5, p103) It was not uncommon, for example, that "AND NOT" would be used instead of NOT or blank would be interpreted as AND. If we could not find any information on whether and how Boolean operators were supported by the search system, we utilized the most frequently used syntax - OR, AND, and NOT - to test these systems. In addition, to test the Boolean operators individually, we added two comparative tests (criterion 15) and evaluated whether queries with different Boolean operators retrieved a valid number of hits. Specifically, we tested first whether the number of hits for "research" minus the number of hits for "research AND define" equalled the number of hits for "research NOT define." Second, we tested whether the number of hits for "research OR define" minus the number of hits for "define" equalled the number of hits for research NOT define. If both tests were passed, we considered Boolean operators functional. Accordingly, the functioning of all three Boolean operators were considered necessary for systematic reviews.

*Criterion 16: "Literal vs Expanded Queries"* determined whether a search system automatically expands

queries impacting on precision and recall. We tested different correct and incorrect versions of the word define to check whether the search system would use autocorrect or would expand the query to different word forms. Additionally, we compared the number of results for terms with different spellings for British and American English. If the number of records for different spellings was the same, we assumed automatic query expansion. As knowledgeable reviewers are able to circumvent automatic query expansion via the use of quotation marks, this criterion was considered desired.

*Criterion 17: "Truncation/Wildcards"* determined whether different frequently used truncation or wildcard symbols were functional. For this criterion, we tested whether terms with truncation or wildcards resulted in more search results than terms without the use of truncation and wildcards. If words with truncation and wildcards produced more hits, they were assumed to function. Similar to criterion 16, the knowledgeable reviewer might circumvent the absence of functional truncation or wildcards by incorporating diverse word forms manually into a search string. Hence, this criterion was considered desired.

*Criterion 18: "Exact Phrase Search"* determined whether the use of quotation marks—symbols typically used to deem an expression should be searched literally—would result in fewer results than for terms lacking them. This is an important feature that allows reviewers to specify exact meanings. In reviewing systematic searches (Table 3), we found that exact phrase searching was used by all systematic reviews. Hence, we deem this criterion necessary.

*Criterion 19: "Parentheses"* determined whether the parentheses functioned in compiling search strings. To create comprehensive search strings with high recall and precision, it is vital to rely on the use of parentheses as these symbols allow a user to group individual concepts and to link them logically. The quick review of systematic searches (Table 3) showed that all systematic reviews used parentheses in their searches. Accordingly, we consider functional parentheses a necessary criterion.

*Criterion 20: "Filtering: Post-Query Refinement"* determined a search system's capacity for post-query refinement through a so-called faceted search. We listed the different filters available to users to refine their search results sets to increase the precision of their search after a query was computed. The more powerful the post-query filter options are, the greater the potential precision of a given query. If search queries work flawlessly and offer options to comprehensively determine search scope, post-query filtering should not be necessary. Hence, we rated post-query refinement of search results as a desired criterion as it is helpful to further specify search scope.

*Criterion 21: "Forward Citation Search"* determined a search system's capacity for forward citation search, that is, the system listing records that cite a specific records of interest. The logic is that records that cite a relevant record will be relevant themselves. Through association, the set of relevant search results can be increased beyond what could have been found through query alone. We reviewed whether forward citation information was offered by a search system, yet did not check the quality of the forward citation search, as this depends on the capacity of the citation index. As the forward citation search is considered a supplementary search method to search queries, we consider it a desired criterion. We did not check search systems' capacity for handsearching in terms of backward citation search, or the search of specific issues or journals.

*Criterion 22: "Advanced Search String Input Field"* was assessed through a review of the search interface. An advanced search input field would allow users to more easily compose advanced search strings. As this limitation can be circumvented with the basic search interface offered by the search system, this criterion is only desired.

*Criterion 23: "Search Help"* was assessed through reviewing the search interface to determine whether the search system provided some documented form of search help to assist users in formulating their search strategies. Search help would primarily mean guidance on which search operators or field codes are available and how users can use the search interface effectively. We consider this criterion to be desired for systematic search.

*Criterion 24: "Maximum Number of Accessible Hits"* was assessed by determining the maximum number of hits made accessible by the search system with a single search. This test required navigating to the last search results page of a query with millions of results to determine how many results were accessible to the reviewer. While hit counts may sometimes run into the millions, it is only a relatively small fraction of this theoretical results set that is actually retrievable in practice. Hence, if a reviewer is interested in the full results set and the hit count goes beyond a set threshold, it is impossible or at least cumbersome to retrieve the full set. One workaround might be to fine-slice the query into smaller results sets that lie below this maximum that can then be handled sequentially. Nevertheless, while this procedure requires significant effort, it is far from certain if it is supported by the specific search system to compile such precise search strings with Boolean operators. We introduced the number of accessible hits as a necessary

criterion. If the criterion was above 1000 records, the performance of the search system was considered sufficient for systematic reviews, yet still not ideal. While searches of with more than 1000 results are common as indicated by our review of systematic searches (Table 3), we opted for a rather conservative threshold as reviewers might mitigate this limitation by dividing search strings to retrieve multiple result sets comprising fewer than 1000 records.

*Criterion 25: "Bulk Download Supported"* was assessed through reviewing the search interface and attempting to download large quantities of results at once. A major time constraint in retrieving search results for subsequent synthesis is search systems' requiring reviewers to download search results in small batches instead of offering full download capabilities. While search system providers want to protect their data from theft and therefore do not provide bulk downloads, this constitutes a tremendous time constraint for reviewers in evidence synthesis and limits their resources for other review activities. We consider this criterion to be desired for systematic search.

*Criterion 26: "Reproducibility of Search Results at Different Times"* tested whether these search queries show signs of bias.[5, p112,54, p141] These tests show whether queries could be repeated so that identical queries retrieve identical results sets, a criterion also described as "the extent to which the search engine returns, from our query, similar results under consistent conditions."[20, p945] Reproducibility is a quality central to systematic reviews that qualifies a search system capable of retrieving results independent of time and place. Accordingly, it is necessary for any search system to offer reproducible results across time in order to meet the quality guidance for systematic searches.

*Criterion 27: "Reproducibility of Search Results at Different Locations"* assessed whether changes in the place from which searches were performed influenced the search results. The place was changed in two forms: we used VPN services to simulate foreign IP addresses and repeated the same query and we logged onto the search system and repeated queries with different institutional access schemes. If these variations produced changes in the search results that could not be explained by the periodic database growth of search systems or by minor discrepancies in the database based on institutional subscription (eg, with Web of Science Core Collections made up of multiple indices that can differ across institutions), the services had to be considered biased. Similar to criterion 26, it is necessary for any search system to offer reproducible results across place in order to meet the quality guidance for systematic searches. A systematic search was considered reproducible when it passed both tests for criteria 26 and 27.

### 2.6 | Principal vs supplementary resources

In our evaluation, a search system could be either rated suitable as a *principal* or *supplementary* resource. A principal resource needed to meet all *necessary* quality requirements or was otherwise considered supplementary. Supplementary resources could be used *in addition to* a principal resource for its specific qualities that could retrieve additional records and to further improve the evidence base. Hence, if a system failed a test for one search method, it might be still considered a good choice for some other search type—for example, while Google Scholar is considered unsuitable for primary review searches, it is considered a suitable supplementary source of evidence (including on grey literature).[57] The distinction between principal and supplementary resources for systematic reviews was also used in previous assessments of search system qualities.[33,57]

Desired quality requirements were not taken into account in this rating of principal and supplementary resources. Instead, these criteria were included in our tests to inform reviewers about functionalities of search systems that were useful or important but still not entirely necessary to meet quality requirements of systematic reviews. The more reviewers are aware of the specific functionalities of a search system, the better they can optimize their search strategy. Accordingly, the tests performed in this study evaluated single search functionalities (see Table 2) so that reviewers received a granular view of how search systems perform for each test. This way, reviewers can quickly consult individual quality criteria for detailed evaluations and reflect on whether a search system of interest offers a service suitable for their needs.

## 3 | RESULTS

Our analysis assessed the suitability of 28 search systems for systematic reviews with 27 test criteria. Each of these criteria assessed a single functionality of the search system. Jointly, these tests showed to what degree a search system was capable of searching effectively and efficiently: qualities necessary for evidence synthesis in the form of systematic reviews. The systematic assessment with these performance tests showed substantial differences in functionality among search systems (see Table 4). While some search systems could be recommended almost without limitation, others failed important tests limiting their suitability for systematic reviews. In other words, not all search systems allow reviewers to perform queries, apply filters, or undertake

citation searching with the high standards required in systematic reviews. Our work makes it possible to classify search systems transparently and objectively according to their suitability for systematic evidence synthesis. We describe the results of these tests that answer questions of interest to a reviewer engaging in systematic search as follows:

1. What is the coverage of the search system, to ensure I access a database suitable for my review?
2. How effectively can I articulate my search via queries, filters, or citation searches so I can retrieve results with high recall and precision?
3. Can I reproduce my search, so that repeated queries will retrieve the same results?
4. How efficiently can I search the system, so I can perform the review within my resource limits?

## 3.1 | Coverage

Our coverage tests assessed five desired criteria a reviewer must consider when choosing a suitable search system. We found that there are significant differences in the databases across all tested criteria. Of the 34 databases offered by the 28 search systems, we found that 16 had a multidisciplinary focus while the remainder were specialized, that is, with a focus on medicine, health sciences, sports, computer science, education, economics, electrical engineering and electronics, psychology, business and management, biomedicine, and transportation studies. The sizes of databases indexed ranged from more than 300 000 to almost 400 million records. Similarly, retrospective coverage ranged between 1550 and 1999, while it is important to note that the number of publications dating back as far as 1550 was small. Further, the number of available record types varied between two and 81 different records. Our sample examined five Open Access resources ("open"), six search systems that focused on Open Access literature, while also providing proprietary resources ("mixed"), and 17 search systems that almost exclusively focused on proprietary content ("proprietary").

## 3.2 | Search queries

We tested the most frequent Boolean operators OR, AND, and NOT via incrementally extended strings adding up to a maximum of six terms and also tested whether Boolean search worked if used with parentheses. Additionally, we assessed exhaustive OR-combinations of different lengths to verify if longer strings also produced more hits. If Boolean operators would not work for short ones, we were not surprised if they also did not work for long ones. Overall, the tests revealed that 17 of the 28 search systems support Boolean operators flawlessly. The remaining search systems retrieved implausible results for search strings consisting of one or more types of Boolean operators. Particularly, AMiner, DBLP, Google Scholar, Microsoft Academic, and WorldWideScience failed all or all but one of the Boolean tests we performed. Further, we found that ERIC seemed to support Boolean searches with our tests of up to six keywords yet failed with longer Boolean searches. The maximum length of the search queries handled without timeouts varies considerably among search systems from only some seven terms (JSTOR) to more than 1000 (EbscoHost, OVID, PubMed, Scopus, Virtual Health Library, Web of Science, and WorldWideScience). For some search systems, we identified restrictions concerning maximum search string length measured in characters, as for example, Google Scholar only allows searches of up to 256 characters. However, for most search strings, the length is determined by the load it puts on the server and thus fails to deliver search results if the request results in a server timeout. The server load seems not only to be determined by search string length but is especially influenced by search scope determined by field codes (eg, title, abstract, or full text search) and the size of the underlying database (searching one or multiple databases via a platform provider).

Of the 28 search systems, 16 seemed to use a form of automatic query expansion to interpret what the user meant instead of processing search strings verbatim. Whenever the default setting is such that queries are expanded automatically, the user can mitigate the effect by adding explicit limiters (in most cases "") to search for keywords literally. On the contrary, explicit limiters ("") were not working or not working correctly in seven search systems. In the cases of AMiner, CiteSeerX, and WorldWideScience, this was especially problematic as these systems expand queries automatically and do not seem to support explicit limiters, forcing reviewers to search via automatically expanded queries. However, for most (21) of the tested search systems, the limiters functioned correctly. We found that while all systems supported the English language, some failed to support Chinese or Cyrillic characters. This may be due to the fact that most texts indexed on these databases were written in English and hence there is little necessity to support additional non-Latin characters. Our tests showed that no search system provided all forms of wildcards and truncation we tested for. Reviewers must exercise caution and test whether the specific truncation or wildcards they use work appropriately.

**TABLE 4** Assessment of 28 academic search systems on their suitability for evidence synthesis

| Name of Search System | Database(s) Searched; Search Settings | 1) Subject | 2) Size | 3)Record Type (Selectable Separately) | 4) Retrospective Coverage (Oldest Entries) | 5) Open Access Content? | 6) Controlled Vocabulary? | 7) Field codes/ Limiters? | 8) Full Text Search Option? | 9) Search String Length | 10) Server Resp. Time/ Records: Max. Word Comb. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | D | D | D | N ≥ 5 | D | N ≥ 25 | N |
| ACM Digital Library | Full index: Full-text collection | Computer science | 2,000,000+ | 4 | 1947 | Proprietary | √ | 17 | √ | 100 | √ |
| AMiner | Full index | Multidisciplinary with a focus on computer science | 233,127,915 | 2 | Unknown | Mixed | X | 3 | X | 10 | X |
| arXiv | Full index; settings: All fields | Multidisciplinary | 1,518,677 | 2 | 1991 | Open | √ | 17 | √ | 100 | √ |
| Bielefeld Academic Search Engine (BASE) | Full index | Multidisciplinary | 144,252,584 | 10 | Unknown | Mixed | √ | 12 | √ | 100 ≤ 1024 characters | √ |
| CiteSeerX | Full index | Multidisciplinary with focus a on computer and information science | 8,401,126 | 3 | Unknown | Open | X | 9 | X | 500 | X |
| ClinicalTrials.gov | Full index | Medicine | 301,373 | 4 | 1999 | Open | √ | 23 | X | 37 | √ |
| Cochrane Library | Cochrane Central Register of Controlled Trials (CENTRAL) | Medicine | 1,317,434 | 5 | 1908 | Proprietary | √ | 15 | √ | 100 | √ |
| Digital Bibliography & Library Project (DBLP) | Full index | Computer science | 4,539,285 | 9 | 1936 | Open | X | 6 | X | 10 | X |
| Directory of Open Access Journals (DOAJ) | Full index | Multidisciplinary | 3,902,698 | 2 | 1874 | Open | √ | 13 | X | 100 | √ |

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 1) Subject | 2) Size | 3)Record Type (Selectable Separately) | 4) Retrospective Coverage (Oldest Entries) | 5) Open Access Content? | 6) Controlled Vocabulary? | 7) Field codes/ Limiters? | 8) Full Text Search Option? | 9) Search String Length | 10) Server Resp. Time/ Records: Max. Word Comb. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | D | D | D | N ≥ 5 | D | N ≥ 25 | N |
| EbscoHost | Selection: ERIC; Medline; EconLit | ERIC: Education studies; Medline: Health studies; EconLit: Economics | ERIC: 1,730,508 Medline: 29,456,831 EconLit: 1,661,780 | 12 | ERIC: 1907 Medline: 1946 EconLit: 1886 | Proprietary | √ | 11 | X | 50 | √ |
| EbscoHost | Selection: CINAHL Plus | Health studies | 6,304,949 | 7 | 1937 | Proprietary | √ | 48 | X | 500 | √ |
| EbscoHost | Selection: SPORTDiscus | Sports studies | 2,449,690 | 6 | 1800 | Proprietary | √ | 27 | X | 1,000 | √ |
| Education Resources Information Center (ERIC) | Full index | Education studies | 1,600,000+ | 24 | 1907 | Proprietary | √ | 19 | √ | 100 | X |
| Google Scholar | Full index | Multidisciplinary | 389,000,000 + | 3 | 1700 | Mixed | X | 5 | √ | 25 ≤ 256 characters | X |
| IEEE Xplore | Full index | Computer science, electrical engineering, electronics | 4,831,568 | 7 | 1872 | Proprietary | √ | 29 | √ | 10 ≤ 15 search terms (can be longer than stated) | √ |
| JSTOR | Full index | Multidisciplinary | 12,000,000+ | 4 | 1857 | Proprietary | √ | 12 | X | 7 (via max. search boxes) | √ |
| Microsoft Academic | Full index | Multidisciplinary | 213,850,455 | 6 | Unknown | Mixed | X | 0 | Unknown | 10 | X |
| OVID | Selection: Embase, Embase Classic | Health studies | 30,000,000+ | 9 | 1947 | Proprietary | √ | 123 | X | 500 | √ |
| OVID | | | 4,000,000+ | 8 | 1806 | Proprietary | √ | 97 | X | 1,000 | √ |

(Continues)

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 1) Subject | 2) Size | 3)Record Type (Selectable Separately) | 4) Retrospective Coverage (Oldest Entries) | 5) Open Access Content? | 6) Controlled Vocabulary? | 7) Field codes/ Limiters? | 8) Full Text Search Option? | 9) Search String Length | 10) Server Resp. Time/ Records: Max. Word Comb. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | D | D | D | N ≥ 5 | D | N ≥ 25 | N |
| | Selection: PsycINFO | Psychological, social, behavioral, health sciences | | | | | | | | | |
| ProQuest | Selection: ABI/INFORM Global | Business, management | 24,190,026 | 13 | 1855 | Proprietary | √ | 29 | √ | 50 | √ |
| ProQuest | Selection: Nursing & Allied Health Database; Public Health Database | Nursing & Allied Health Database: Health studies; Public Health Database: Health studies | 12,954,446 | 14 | Nursing & Allied Health Database: 1857; Public Health Database: 1934 | Proprietary | √ | 34 | √ | 25 | √ |
| PubMed | Full index: Medline (and others) | Biomedicine; health studies | 29,000,000+ | 3 | 1790 | Mixed | √ | 29 | X | 1,000 | √ |
| ScienceDirect | Full index | Multidisciplinary | 15,000,000+ | 24 | 1823 | Proprietary | X | 14 | X | 25 ≤ 8 Boolean connectors allowed (actually more possible) | √ |
| Scopus | Full index | Multidisciplinary | 70,000,000+ | 13 | 1861 | Proprietary | √ | 79 | X | 1,000 | √ |
| Semantic Scholar | Full index | Multidisciplinary with a focus on computer science and medicine | 72,366,665 | 10 | 1931 (or earlier) | Proprietary | X | 0 | Unknown | 500 | X |
| Springer Link | Full index | Multidisciplinary | 12,731,539 | 10 | 1815 (single documents) | Proprietary | √ | 4 | X | 50 | √ |

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 1) Subject | 2) Size | 3)Record Type (Selectable Separately) | 4) Retrospective Coverage (Oldest Entries) | 5) Open Access Content? | 6) Controlled Vocabulary? | 7) Field codes/ Limiters? | 8) Full Text Search Option? | 9) Search String Length | 10) Server Resp. Time/ Records: Max. Word Comb. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | D | D | D | N ≥ 5 | D | N ≥ 25 | N |
| Transport Research International Documentation (TRID) | Full index | Transportation studies | 1,200,000+ | 5 | 1900 | Proprietary | √ | 14 | X | 100 | √ |
| Virtual Health Library | Full index | Health studies | 865,836 | 10 | 1902 (single studies); 1966 full | Proprietary | √ | 13 | X | 1,000 | √ |
| Web of Science | Selection: Web of Science Core Collection[a] | Multidisciplinary (depends on selected databases) | 73,000,000+ | 21 | 1900 (depends on underlying subscription) | Proprietary | √ | 18 | X | 1,000 | √ |
| Web of Science | Selection: Medline | Health studies | 29,303,305 | 81 | 1950 | Proprietary | √ | 24 | X | 1,000 | √ |
| Wiley Online Library | Full index | Multidisciplinary | 8,000,000+ | 3 | 1798 | Proprietary | √ | 7 | X | 100 | √ |
| WorldCat | Selection: Thesis/ dissertation | Multidisciplinary | 8,000,000+ | 17 | About 1550 for earliest theses | Proprietary | √ | 13 | X | 25 | √ |
| WorldWideScience | Full index | Multidisciplinary | 323,000,000 | 17 | 1869 (single studies) | Mixed | √ | 6 | √ | 1,000 | X |

**TABLE 4** Assessment of 28 academic search systems on their suitability for evidence synthesis

| Name of Search System | Database(s) Searched; Search Settings | 11) Language | 12) Boolean Functional? OR | 13) Boolean Functional? AND | 14) Boolean Functional? NOT | 15) Comparative Test | 16) Query Interpretation/ Query Expansion | 17) Truncation/ Wildcards Available? | 18) Exact Phrases Functional? | 19) Parenthesis Functional? | 20) Post-query Results Refinement |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | N | N | N | N | D | D | N | N | D |
| ACM Digital Library | Full index: Full-text collection | √ (E, Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 9 |
| AMiner | Full index | √ (E, Ch, Cy) | X | X | X | X | X | X | X | X | 0 |
| arXiv | Full index; settings: All fields | √ (E), X (Ch, Cy) | √ | √ | √ | X | √ | X | √ | X | 0 |
| Bielefeld Academic Search Engine (BASE) | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 9 |
| CiteSeerX | Full index | √ (E, Ch, Cy) | X | √ | √ | X | X | X | X | X | 0 |
| ClinicalTrials.gov | Full index | √ (E, Cy) X (Ch) | √ | √ | √ | √ | X | X | √ | √ | 12 |
| Cochrane Library | Cochrane Central Register of Controlled Trials (CENTRAL) | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 3 |
| Digital Bibliography & Library Project (DBLP) | Full index | √ (E, Cy) X (Ch) | X | X | X | X | √ | X | X | X | 4 |
| Directory of Open Access Journals (DOAJ) | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | √ | X | X | X | 6 |
| EbscoHost | Selection: ERIC; Medline; EconLit | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 13 |
| EbscoHost | Selection: CINAHL Plus | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 12 |

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 11) Language | 12) Boolean Functional? OR | 13) Boolean Functional? AND | 14) Boolean Functional? NOT | 15) Comparative Test | 16) Query Interpretation/ Query Expansion | 17) Truncation/ Wildcards Available? | 18) Exact Phrases Functional? | 19) Parenthesis Functional? | 20) Post-query Results Refinement |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | N | N | N | N | D | D | N | N | D |
| EbscoHost | Selection: SPORTDiscus | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 12 |
| Education Resources Information Center (ERIC) | Full index | √ (E) X (Ch, Cy) | √ | √ | √ | √ | √ | X | X | √ | 11 |
| Google Scholar | Full index | √ (E, Ch, Cy) | X | X | √ | X | X | X | √ | X | 2 |
| IEEE Xplore | Full index | √ (E) X (Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 12 |
| JSTOR | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 4 |
| Microsoft Academic | Full index | √ (E, Ch, Cy) | X | X | X | X | √ | X | X | X | 7 |
| OVID | Selection: Embase, Embase Classic | √ (E) X (Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 5 |
| OVID | Selection: PsycINFO | √ (E) X (Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 6 |
| ProQuest | Selection: ABI/INFORM Global | √ (E, Ch, Cy) | √ | √ | √ (with adapted search string) | √ | X | X | √ | √ | 10 |
| ProQuest | Selection: Nursing & Allied Health Database; Public Health Database | √ (E, Ch, Cy) | √ | √ | √ (with adapted search string) | √ | X | X | √ | √ | 13 |
| PubMed | Full index: Medline (and others) | √ (E, Cy) X (Ch) | √ | √ | √ | √ | X | X | √ | √ | 10 |

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 11) Language | 12) Boolean Functional? OR | 13) Boolean Functional? AND | 14) Boolean Functional? NOT | 15) Comparative Test | 16) Query Interpretation/ Query Expansion | 17) Truncation/ Wildcards Available? | 18) Exact Phrases Functional? | 19) Parenthesis Functional? | 20) Post-query Results Refinement |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | N | N | N | N | D | D | N | N | D |
| ScienceDirect | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 4 |
| Scopus | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 11 |
| Semantic Scholar | Full index | √ (E, Ch, Cy) | X | √ | √ | X | √ | X | √ | X | 4 |
| Springer Link | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 6 |
| Transport Research International Documentation (TRID) | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 14 |
| Virtual Health Library | Full index | √ (E, Ch, Cy) | √ | √ | √ | √ | √ | X | √ | √ | 13 |
| Web of Science | Selection: Web of Science Core Collection[a] | √ (E) X (Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 18 |
| Web of Science | Selection: Medline | √ (E, Ch, Cy) | √ | √ | √ | √ | X | X | √ | √ | 13 |
| Wiley Online Library | Full index | √ (E, Ch) X (Cy) | √ | √ | √ | √ | X | X | √ | √ | 6 |
| WorldCat | Selection: Thesis/ dissertation | √ (E, Ch, Cy) | √ | √ | √ | X | X | X | √ | √ | 7 |
| WorldWideScience | Full index | √ (E, Ch, Cy) | X | X | X | X | X | X | X | X | 12 |

**TABLE 4** Assessment of 28 academic search systems on their suitability for evidence synthesis

| Name of Search System | Database(s) Searched; Search Settings | 21) Citation Search (Forward) | 22) Advanced Search String Field? | 23) Search Help? | 24) No. of Accessible Hits | 25) Bulk Download? | 26) Repeatable? Time | 27) Location-Independent? IP | Assessment |
|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | N ≥ 1,000 | D | N | N | |
| ACM Digital Library | Full index: Full-text collection | √ | √ | X | Full | 2,000 | √ | √ | PRINCIPAL |
| AMiner | Full index | √ | √ | X | 1,000 | X | √ | √ | SUPPLEMENTARY |
| arXiv | Full index; settings: All fields | √ | √ | √ | 10,000 | X | √ | √ | SUPPLEMENTARY |
| Bielefeld Academic Search Engine (BASE) | Full index | X | √ | √ | 1,000 | 100 | √ | √ | PRINCIPAL |
| CiteSeerX | Full index | √ | √ | √ | 500 | X | √ | √ | SUPPLEMENTARY |
| ClinicalTrials.gov | Full index | X | √ | √ | Full | Full | √ | √ | PRINCIPAL |
| Cochrane Library | Cochrane Central Register of Controlled Trials (CENTRAL) | X | √ | √ | Full | Full | √ | √ | PRINCIPAL |
| Digital Bibliography & Library Project (DBLP) | Full index | X | X | √ | Unknown (most likely full) | X | √ | √ | SUPPLEMENTARY |
| Directory of Open Access Journals (DOAJ) | Full index | X | √ | X | 10,010 | X | √ | √ | SUPPLEMENTARY |
| EbscoHost | Selection: ERIC; Medline; EconLit | X | √ | √ | 25,000 | 25,000 | √ | √ (depends on database access) | PRINCIPAL |
| EbscoHost | Selection: CINAHL Plus | X | √ | √ | 25,000 | 25,000 | √ | √ (depends on database access) | PRINCIPAL |
| EbscoHost | Selection: SPORTDiscus | X | √ | √ | 25,000 | 25,000 | √ | √ (depends on database access) | PRINCIPAL |

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 21) Citation Search (Forward) | 22) Advanced Search String Field? | 23) Search Help? | 24) No. of Accessible Hits | 25) Bulk Download? | 26) Repeatable? Time | 27) Location-Independent? IP | Assessment |
|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | N ≥ 1,000 | D | N | N | |
| Education Resources Information Center (ERIC) | Full index | X | X | √ | Full | 200 | √ | √ | |
| Google Scholar | Full index | √ | √ | √ | 1,000 | X | X | X | SUPPLEMENTARY |
| IEEE Xplore | Full index | X | √ | √ | 2,000 | 2,000 | √ | √ | SUPPLEMENTARY |
| JSTOR | Full index | X | √ | √ | 1,000 | X | √ | √ | SUPPLEMENTARY |
| Microsoft Academic | Full index | √ | X | X | 5,000 | X | √ | √ | SUPPLEMENTARY |
| OVID | Selection: Embase, Embase Classic | √ | √ | √ | Full | 1,000 | √ | √ (depends on database access) | PRINCIPAL |
| OVID | Selection: PsycINFO | √ | √ | √ | Full | 1,000 | √ | √ (depends on database access) | PRINCIPAL |
| ProQuest | Selection: ABI/INFORM Global | √ | √ | √ | 10,000 | 100 | √ | √ (depends on database access) | PRINCIPAL |
| ProQuest | Selection: Nursing & Allied Health Database; Public Health Database | √ | √ | √ | 10,000 | 100 | √ | √ (depends on database access) | PRINCIPAL |
| PubMed | Full index: Medline (and others) | √ | √ | √ | Full | Full | √ | √ | PRINCIPAL |
| ScienceDirect | Full index | √ | √ | √ | 6,000 | 100 | √ | √ | PRINCIPAL |
| Scopus | Full index | √ | √ | √ | 2,000 | 20,000 | √ | √ | PRINCIPAL |
| Semantic Scholar | Full index | √ | √ | √ | 10,000 | X | √ | √ | SUPPLEMENTARY |
| Springer Link | Full index | X | √ | √ | 19,980 | 1,000 | √ | √ | SUPPLEMENTARY |
| Transport Research International Documentation (TRID) | Full index | X | X | √ | 15,000 | X | √ | √ | PRINCIPAL |
| Virtual Health Library | Full index | X | √ | √ | Full | Full | √ | √ | PRINCIPAL |
| Web of Science | Selection: Web of Science Core Collection[a] | √ | √ | √ | 100,000 | 5,000 | √ | √ (depends on database access) | PRINCIPAL |

**TABLE 4** (Continued)

| Name of Search System | Database(s) Searched; Search Settings | 21) Citation Search (Forward) | 22) Advanced Search String Field? | 23) Search Help? | 24) No. of Accessible Hits | 25) Bulk Download? | 26) Repeatable? Time | 27) Location-Independent? IP | Assessment |
|---|---|---|---|---|---|---|---|---|---|
| | | D | D | D | N ≥ 1,000 | D | N | N | |
| Web of Science | Selection: Medline | √ | √ | √ | 100,000 | 5,000 | √ | √ (depends on database access) | PRINCIPAL |
| Wiley Online Library | Full index | √ | √ | √ | 2,000 | X | √ | √ | PRINCIPAL |
| WorldCat | Selection: Thesis/ dissertation | X | √ | √ | 5,000 | X | √ | √ | SUPPLEMENTARY |
| WorldWideScience | Full index | X | √ | √ | Limited (only top results are shown) | 20 | X | X | SUPPLEMENTARY |

*Note.* √, passed test; X, failed test;
Abbreviations: D, desired; N, necessary.
[a]Science Citation Index Expanded (1900-present); Social Sciences Citation Index (1900-present); Arts & Humanities Citation Index (1975-present); Conference Proceedings Citation Index- Science (1990-present); Conference Proceedings Citation Index- Social Science & Humanities (1990-present); Book Citation Index– Science (2010-present); Book Citation Index– Social Sciences & Humanities (2010-present); Emerging Sources Citation Index (2015-present); Current Chemical Reactions (2010-present); (Includes Institut National de la Propriete Industrielle structure data back to 1840); Index Chemicus (2010-present).

We found that all but four search systems (24) included some form of advanced search field where users could structure their systematic searches. While some systems possessed multiple forms with varying degrees of complexity according to the needs and search literacy of the users (eg, ProQuest and Web of Science), others only provided rudimentary interfaces, leaving the user little freedom to specify requests (eg, AMiner and Microsoft Academic). All systems except Microsoft Academic and Semantic Scholar had some form of field codes that allowed the user to specify which parts of the structured information available in the databases to access. Both these search systems relied on semantic searching, where, in contrast to a traditional literal search, the focus is on interpreting what users might have *meant*, instead of retrieving results according to exactly what the search terms specified. Following a semantic search request, the search results can then be customized with a limited number of post-query filters. In contrast, some traditional search systems, like PubMed, and IEEE Xplore offer close to 30 different options to filter their highly structured databases, which is especially convenient for the structured queries necessary in the fields of medicine and engineering. For platforms, the number and type of field codes depends on the underlying database. All but four search systems (24) offered some form of search help to assist users in conducting their search.

While most (21 of 28) search systems offer some kind of controlled vocabulary, the quality differs significantly. In medicine, reviewers rely on frequently updated and rigorously categorized Medical Subject Headings (MeSH), while in other disciplines, specialist databases offer simpler thesauri. The availability of controlled vocabulary depends on the underlying database and its data structure. We found that in 79% of the cases, the controlled vocabulary was presented in a hierarchical form with multiple levels, and in 61% of the cases, the controlled vocabulary was searchable. Full text search functionality was available in only nine search systems, while 17 search systems did not provide that functionality, and for the semantic search engines Microsoft Academic and Semantic Scholar, it was unclear what parts of the records are indexed and searched.

While AMiner, arXiv, and CiteSeerX do not provide any post-query refinements, other systems such as EbscoHost, ProQuest, and Web of Science provide up to 18 different options for filtering content. However, these refinement options depend significantly on the reviewers' selection of databases searched. As the databases hosted on these systems differ in their structured information, the options for filtering them differ

accordingly. Forward citation searching was available on more than half (15) of all search systems.

### 3.3 | Search results

We found that the maximum number of retrievable records varies greatly among search systems. While some allow access to all records, the functionality of CiteSeerX and WorldWideScience is severely restricted as their threshold lies below 1000 records. Most notably, ACM Digital Library, ClinicalTrials.gov, Cochrane Library, ERIC, OVID, PubMed, and Virtual Health Library allow full access to all datasets returned from a single search. For DBLP, we could not determine the scope of retrievable records since its results set expands dynamically rather than using numeration or pagination. Similarly, bulk download options differed significantly across search systems. Some allow the download of the entire search results set in one go, while almost half of the examined systems provided no support for for exporting multiple records. Most positively, the medical databases of Virtual Health Library, ClinicalTrials.gov, and Cochrane Library supported efficient data retrieval by allowing full download options. While many databases offered an application programming interface (API) to access their database, these options are only accessible to reviewers with programming skills.

### 3.4 | Search reproducibility

In our sample of 28 academic search systems, all but two—Google Scholar and WorldWideScience—were reproducible in terms of reporting identical results for repeated identical queries. While WorldWideScience failed to deliver replicable results at all times, Google Scholar failed to deliver them only during certain periods: sometimes, search results were replicable with two consecutive queries; then with a third query or with queries after some queries in between, they were no longer replicable and the results set differed in a way not explainable by *natural* database growth. Natural growth means that the dataset indexed on a database increases with the identification and curation of new records and thus that results sets retrieved tend to increase with repeated queries as the underlying database has expanded in the meantime. All the other 26 search systems appeared to provide reproducible results.

Further, in our analysis of search results retrieved via a changed retrieval location, we found differences for varying institutional subscriptions, yet not for differences in IP addresses. Certain subscription-based platforms—for example, EbscoHost, OVID, ProQuest, and Web of Science—delivered notably different results depending on the institution through which we accessed the underlying databases. For these systems, the number of records depended on the subscriptions to different databases or indexes subscribed to by the organization. In most cases, differences depended on the databases available, yet there were also differences within the same databases. We found the coverage of the same database was different as the number of years accessible varies from package to package. These differences can be visible in the description of the database, as with ProQuest offering its popular ABI/Inform package in different versions containing substantially different results sets. Nevertheless, these differences are sometimes not so obvious for users, requiring closer examination. Web of Science's Core Collection also varies significantly in scope containing different indices depending in the subscription. These single indices, again, vary in scope for the subscribing libraries. For the same index, one institution might have subscribed to a retrospective coverage since 1996, another since 2010. The variations highlight that reviewers should be familiar with their institution's subscription and that they need to document this in detail in their review reports.

## 4 | DISCUSSION

Overall, we found that only 14 of the 28 academic search systems examined are well-suited to evidence synthesis in the form of systematic reviews in that they met all necessary performance requirements (Table 4). These 14 can be used as principal search systems: ACM Digital Library, BASE, ClinicalTrials.gov, Cochrane Library, EbscoHost (tested for ERIC, Medline, EconLit, CINHAL Plus, SportsDiscus), OVID (tested for Embase, Embase Classic, PsychINFO), ProQuest (tested for Nursing & Allied Health Database, Public Health Database), PubMed, ScienceDirect, Scopus, TRID, Virtual Health Library, Web of Science (tested for Web of Science Core Collection, Medline), and Wiley Online Library. In contrast, the remaining 14 were unsuitable for use as the principal search system for systematic reviews due to failing to meet one or more necessary criteria. For these 14 search systems, our tests uncovered severe performance limitations with regard to formulating queries, the correct interpretation of queries by the system, data retrieval capabilities, and the reproducibility of searches. These systems should only be considered supplementary to the principal systems, especially for nonquery-based search methods where they might still provide great benefit. Desired

performance criteria inform about other-nonessential functionalities relevant for systematic search, where reviewers need to assess individually how important these criteria are for their specific search. We next present the results of our analysis. The criteria we base our assessment on can be found in Table 2, and the detailed outcomes of the tests conducted can be found in Appendix II (supplementary online material).

## 4.1 | Necessary criteria

In most systematic reviews, Boolean queries retrieve the largest portion of relevant records because they allow the user to search large databases with the highest recall. Accordingly, the query-based search form is the backbone of systematic reviews. It is striking that half the search systems we examined have at least some issues with Boolean queries. This is particularly unfortunate because Boolean searching is effective, especially for systematic search strategies: "medical research indicates that expert searching based on Boolean systems is still the most effective method [of searching]."[26, p1570]

Our tests revealed that the help files of numerous search systems promise a Boolean search functionality that our tests could not verify. These findings were especially alarming because users of such systems rely on functionalities that they assume to work properly, but that may not be the case. In a review of search FAQs, we found that arXiv, CiteSeerX, DBLP, ERIC, Semantic Scholar, WorldCat, and WorldWideScience promote support for Boolean searching, yet our tests identified issues with searches using Boolean operators. Semantic Scholar, for example, writes in its FAQ "you may search for papers on Semantic Scholar using AND/OR query terms," yet it failed most query-based tests (criteria 10, 12, 15, and 19). For these services, the negative performance results might hint at glitches that the system administrators should examine. The other systems that failed query tests—AMiner, DOAJ, Google Scholar, and Microsoft Academic—do not state support for Boolean search functionality.

Given the findings in this study, we advocate reassessing the advice given in evidence-synthesis guidance for systematic reviews. For example, the searching guidance of The Campbell Collaboration states: "Given an Internet search engine (Google, Google Scholar, Bing, etc.) [...], many of these search strategies may also be applied. For example, Phrase searching, Boolean Operators and Limiting features are typically all offered. Using the search engine's Advanced search screen can provide an easy way of accessing these features."[15, p34] This passage might incorrectly advise users to pursue full Boolean search strategies with search systems such as Google Scholar that do not offer such functionality. Further, our results contradict systematic review guidance that assumes that "all the search engines in some way [would] permit the use of Boolean syntax operators to expand or restrict the search."[5, p103] The results of our study show that when it comes to search functionalities that are necessary for systematic reviews, a reviewer must look closely at which search systems are in fact suitable and why. If reviewers are comfortable with search systems failing specific *desired* criteria, while query capabilities are sufficient, they should not be discouraged from using it as their principal search system.

## 4.2 | Desired criteria

Search systems failing one or more necessary test criteria are *always* in conflict with the fundamental quality requirements of systematic reviews, especially for searches with search strings. Accordingly, we advise reviewers to use these systems solely for supplementary search methods as they might still be valuable in improving search outcome. Such supplementary methods include handsearching of backward and forward citations, specific issues or journals, or the use of filters to limit search results. We explicitly tested for handsearching in the form of forward citation searching, as this information needs to be provided by a citation index that contains information on which records have cited a specific record. This citation index is, however, not available through every search system. While our methods did not allow the testing of the comprehensiveness of citation indexes, larger, multidisciplinary search systems seem to provide more complete citation information than smaller specialized search systems. Comparisons of citation indexes generally rate Google Scholar as the most comprehensive.[71-73] These comparisons support the idea that larger search systems tend to have greater citation coverage. While our results show that 15 of the 28 systems examined have cross-citation information, because of the limited coverage of many of these systems', their citation information might be limited as well. Hence, if the reviewer's goal is to reach beyond the limitations of a specialized search system, it might prove beneficial to use citation information from a large, multidisciplinary search system to broaden the search scope.

Desired performance criteria need to be evaluated relative to the *specific* systematic search requirements of the reviewer. Our analysis made some important evaluation criteria transparent, so it is possible for reviewers to reflect on how these criteria could facilitate or limit their

systematic searches. For reviewers, it is important to choose a search system that is suitable for a given research domain, a certain retrospective focus, and that covers the specific record type of interest. Coverage of a search system and/or its underlying database(s) might be important for evaluating a search system's potential recall. Coverage is, however, only beneficial when the necessary retrieval capabilities are offered as well. Otherwise, searching large, multidisciplinary databases might involve low search precision, making systematic search inefficient and laborious. Alternatively, reviewers might want to test systems offering the option to download resources in bulk. Compared with systems without this feature, bulk download allows efficient data handling in combination with reference management software and data analysis tools.

Our analysis showed that of the five Open Access search systems examined that catalogue Open Access records only (arXiv, CiteSeerX, ClinicalTrials.gov, DBLP and DOAJ), only ClinicalTrials.gov passed all tests relating to necessary criteria. Among the six systems offering mixed access, that is, using a dataset offering both proprietary and Open Access content, BASE and PubMed were found to be suitable for use as principal search systems. Accordingly, for reviewers having no access to proprietary databases, our findings mean they only have a limited selection of Open Access database alternatives. Reviewers interested in medical evidence synthesis could access all three—BASE, PubMed, and ClinicalTrials.gov—but should be aware that BASE also indexes PubMed. Reviewers from other disciplines who want synthesise Open Access content systematically, however, are limited to the multidisciplinary system BASE, which provides full texts for 60% of its close to 150 million records under Open Access licence. Other open, or partially open, search systems that fail to meet the criteria for query-based search might still be useful for supplementary search methods.

### 4.3 | Differences in search systems

The tests applied in this study not only compared individual search systems but also underlying databases accessed through different platforms. For example, ERIC's database is accessible via its dedicated search system, but also through EbscoHost. Similarly, we accessed Medline through EbscoHost, PubMed, and Web of Science (and indirectly through BASE and other systems that use the Medline index). The analysis detailed above detected some performance differences. While the underlying database seemed largely identical, determined by its size, the functionalities of the search system through which it was accessed varied. ERIC (the search system), for example, failed some necessary tests, whereas when accessed through EbscoHost, the search functionalities in searching the ERIC database were superior. We also identified differences in the case of Medline. While PubMed allows bulk download of the full dataset, EbscoHost allows 25 000 and Web of Science 5000. Hence, we conclude that in these cases the search capabilities depended on the system through which it was accessed and less on the underlying database.

Additionally, the platforms in our sample—EbscoHost, OVID, ProQuest, and Web of Science—all underwent multiple tests where individual platforms were tested with varying databases. These repeated tests were aimed to provide another perspective on performance determinants of platforms and should show whether changing underlying databases influenced necessary and desired performance criteria. As the underlying databases changed, so accordingly did the results for tests of these databases, such as scope, available record types, controlled vocabulary, retrospective coverage, field codes, or filters. Further, we found that the maximum length of the search string a platform could handle without timeout differed significantly depending on the size and number of underlying databases. The scope of the search determined by field codes also seems to influence server load and thus maximum search string length. This means a full text search puts a heavier load on the system than, for example, a search of titles or abstracts. Hence, it is not the number of characters (alone) that determines the longest still computable search string. Reviewers may thus need to balance search string length with database selection and field code selection in the case of more exhaustive searches. Another option might be to split the string into pieces and search systems sequentially, although doing so would extend the workload associated with documentation and deduplication.

The quality of systematic searches not only depends on the queries or filters specified for searching a database, but also depends on the database itself. Database providers/platforms, such as EbscoHost, OVID, ProQuest, and Web of Science, provide access to multiple databases simultaneously. Hence, for these platforms, reviewers need to report on the exact databases they have searched. Nevertheless, inexperienced researchers frequently wrongly assume they are searching a single, distinct database, while in truth, that search system aggregates multiple databases. The consequence is that these authors report using a search system but omit to record using its underlying databases. With this limited information, the search process is insufficiently documented and replication is impossible. Hence, for systematic searches of platforms such as EbscoHost, OVID, ProQuest, and Web of

Science, we remind authors that they must report the underlying databases and the indices they contain.[74] Further, it is necessary to bear in mind that databases update frequently—sometimes multiple times a day or even on an hourly basis—thus the underlying dataset changes accordingly. While most of the time the dataset will increase through the addition of records, databases can also shrink through the deletion of duplicates or the occurrence of errors that affect the dataset provided. It is therefore essential to report—in addition to the exact database accessed—the time the dataset was accessed too. One option to facilitate these reporting practices might be to include such requirements in the "guide for authors" section of journals and to advise the use of reporting guidance such as PRISMA[36] or ROSES.[74]

## 4.4 | Emergence of semantic search systems

There has recently been an upsurge in using semantic search engines over traditional ones, as is evident in the birth of Semantic Scholar (2015), the relaunch of Microsoft Academic (2017), and the expected launch of *Meta*, a project of the Zuckerberg foundation. These semantic search engines tend to be designed to reward exploratory rather than systematic search behavior. These tendencies add to the notion that "the problem is [...] that an ideological tendency to make things 'user friendly' (and the market bigger) tends to hurt the development of systems aimed at increasing the selection power of users and search experts."(26, p1570) Our findings indicating that these systems are inadequate to be used as principal systems in systematic searches support this notion. The criticism of user-friendliness at any cost is especially directed at Google Scholar, which is more concerned with "*tuning*" its first results page(75, p15) than with overall precision. This makes Google Scholar highly precise for exploratory searches conducted by a user interested in only a few relevant results on the first search engine results page.[76,77] Nevertheless, overall, Google Scholar's search precision has been found to be significantly lower than 1% for systematic searches.[25] This is not surprising, since our findings show that Google Scholar does not support many of the features required for systematic searches. Our findings support the criticism of Bramer et al,[33] Bramer et al,[34] and Boeker et al[25] and indicate that Google Scholar's coverage and recall is an inadequate reason to use it as principal search system in systematic searches.[53] If a system such as Google Scholar fails to deliver retrieval capabilities that allow a reviewer to search systematically with high levels of recall, precision, transparency, and reproducibility, its coverage is irrelevant for query-based search. Google Scholar's extraordinary coverage acting as a multidisciplinary compendium of scientific world knowledge should not blind users to the fact that users' ability to access this compendium is severely limited, especially in terms of a systematic search.

While popular search systems such as Google Scholar or Microsoft Academic being inadequate for query-based search is already unfortunate on its own, the situation is made worse by users seemingly being unaware of these shortcomings. Users are perhaps guided by convenience rather than strategic considerations when choosing their search system. In fact, it was due to its great ease of use and performance in informational and exploratory searches[78] that Google Scholar emerged as the number one go-to academic search engine for most academic users.[79-82] Students in particular utilize Google as a main source in information seeking.[52,83,84] The requirements for evidence synthesis are not always obvious to reviewers as they are used to navigational or exploratory searching that comes intuitively.[78,85-87] Students especially seem to have tremendous difficulty in mastering online literature searches.[83,88-91] Other search systems, such as bibliographic databases or platforms, are less popular due to the elevated skills they require and, in some cases, more difficult access due to paywall restrictions. We advocate educated use of these systems, so users have the right tool for the right purpose fully aware of its strengths and weaknesses.

## 4.5 | Limitations

While we took the greatest care to include a large evidence-based selection of meaningful methods to test the capacity of search systems, there may be other tests unknown to us that could be performed. The tests conducted here do, however, rigorously and transparently assess whether and to what extent search systems succeed in such tests compared to other systems. Generally, we did not directly test a search system's *level of* precision or recall, but rather *the capacity of allowing the user to specify queries* with high levels of precision and recall. From a methodological standpoint, this study is particularly influenced by the research of Gusenbauer[37] and Boeker et al(25, p11) that sought to, "to compare the effectiveness of Google Scholar and other retrieval tools" Our approach provides great practical benefit, as it illuminates some of the most critical strengths and weaknesses of search systems that are often only communicated without comparative evidence of actual performance criteria. This study contributes such tangible criteria. Nevertheless, from a theoretical standpoint, our study can only

provide evidence that search systems behave incorrectly in failing to comply with certain test criteria. It is impossible to be absolutely certain that a system that has proved successful in our specific tests would not fail in slightly different tests or under different circumstances. For instance, a system providing a functional 1000-term OR-string could theoretically fail one consisting of 1001 terms or let us say 521 terms. Similarly, it is impossible to rule out that search systems have temporal performance variations that cannot be captured with cross-sectional analysis. While we also included a test for temporal variation through reproducibility tests where we determined search engine bias, we have to assume that the systems are otherwise stable in scenarios that lie beyond our tested scope.

One possible limitation might be that whenever a certain threshold is defined, someone asks why it was not some other threshold. In this study, we tried to alleviate this criticism by basing our thresholds on the quality guidance issues by Cochrane, The Campbell Collaboration, and the CEE. Further, if we needed to decide on specific numeric thresholds such as the minimum length of search strings or the minimum number of field codes, we based our decision on a review of best practices from previously published and highly cited systematic reviews.

As most of the search systems update not only their database, but also their search functionalities, the performance results tested in this study might change over time. However, during the period covered by writing-up the results of this study, those results remained relatively stable, which most likely reflects the fact that our performance tests evaluated fundamental functionalities of search systems that are rarely updated. Nevertheless, to be sure to have an accurate picture of search functionalities, reviewers can easily replicate our tests and evaluate them immediately before they access their search system of interest.

## 5 | CONCLUSION

Selection of suitable search systems is essential for the outcome of evidence-synthesis research. Reviewers must consider the different functionalities offered, or not offered, when interacting with a given search system. In particular, they must be cognisant of the trade-off between search precision and recall. Searching search systems with the greatest effectiveness and efficiency is a skill that is necessary yet generally undervalued in education and research practice. Reviewers should always consult information specialists or librarians and enlist their support in designing systematic review search strategies.[9,13] Only if reviewers are aware of a search system's functionalities, they can take advantage of all methods and functionalities and design good search strategies.

Yet, so often convenience guides the method of search system choice. Unfortunately, awareness of the differences between search systems is not yet sufficiently developed in the area of scientific education. Indeed, librarians affirm the lack of search skills prevalent especially among students[89,92] and so-called digital natives.[88,93] We hope our study helps to create awareness of the importance of search literacy. This study shows the limitations of such convenience. This research encourages responsible and knowledgeable researchers to be aware of search system qualities so they can then use the appropriate tool for the task at hand. Just as artisans have particular tools for particular tasks, we should understand our digital tools are not one-size-fits-all solutions. Crawler-based search engines like Google Scholar or Microsoft Academic function differently to database providers such as ProQuest or EbscoHost, or journal platforms such as SpringerLink or Wiley. The overview provided here should make it easier for scholars to choose the most adequate search system according to their unique information requirements.

The many limitations we identified affecting most of the search systems in our study clearly call for researchers - especially those who engage in systematic searches - to ensure that they possess considerable knowledge of the search systems they intend to use; however, those qualities are not always evident. Without this knowledge, search results might be misinterpreted in a way that impinges on research validity. A high number of hits resulting from an extensive Boolean search string might, for example, be seen as indication of a high number of relevant records—yet in truth, due to faulty interpretation of the search system, might reflect the malfunction of limiting AND or NOT operators.

It has often been unclear exactly why certain search systems perform better or worse than others. Performance issues, especially those concerning the correct interpretation of Boolean search strings by search systems, may have remained undetected so far. We aimed to make these performance differences explicit. Since we used the same metrics for all systems, our assessment makes a large set of systems comparable. If impediments of search systems are made transparent, experienced reviewers could perhaps circumvent these limitations by using search systems differently. However, researchers lacking such knowledge run the risk of expecting too much of search systems (even when searched in a systematic way) and drawing erroneous conclusions based on biased sets of search

results. The establishment of the 27 testing criteria here may help to create awareness among reviewers of where they need to look when selecting and using search systems.

If the results of scientific research are to be cumulative, researchers in general, and especially those aiming to conduct evidence syntheses, should know how to effectively and efficiently gather scientific knowledge. Our evaluation offers reviewers a means of transparent evaluation. While some researchers highlight the benefits of easy-to-use academic search engines like Google Scholar[53] that allow non-experts to make use of scholarly resources,[94] our work highlights the specific pitfalls of those systems. In contrast, we demonstrate that using search systems correctly is not always as straightforward as slick user interfaces might suggest. Further, our detailed assessment based on 27 transparent criteria is also especially helpful for experienced reviewers of all disciplines when they decide on which criteria their search system of choice must meet. The distinction between necessary and desired criteria should create awareness of why certain search systems are suitable and others unsuitable, and that simple distinction could be helpful, especially for non-expert reviewers or those who are not information specialists.

Our analysis reveals that few Open Access search systems can be recommended as a principal resource for systematic searches. It seems there is currently almost no getting around proprietary search systems if one attempts a rigorous systematic review. This finding is extremely unfortunate, as Open Access databases advocate barrier-free access to information, yet for systematic reviews, they most often do not provided the necessary functionalities to be used as principal search systems. For researchers from resource-constrained contexts, we could perhaps recommend the multidisciplinary BASE, as it is a comprehensive resource with a large share of Open Access content and it also met necessary testing criteria in our analysis.

We advocate that search system operators—Open Access or not—review the capabilities and improve performance criteria where necessary. Ideally, search system providers would use our insights to further develop their systems according to the high standards of evidence-synthesis guidance. It becomes evident that these providers need to balance the pros and cons of "exact-match systems" and "best-match systems" or find ways of alleviating the effects of trade-offs between both concepts.[26] The metrics used in this research might prove helpful in defining some of the specifications an improved system should possess—something especially relevant for those systems in which we uncovered severe performance limitations. Such scientific performance requirements might become increasingly relevant with the current research trend of replicating existing studies, and with the continued increase in the number of published systematic reviews.

The criteria established in this study are relatively straightforward as they are defined from the viewpoint of the reviewer. Therefore, it is easily possible for reviewers to update these assessments frequently by identifying changes in the qualities of the single search systems or adding previously unexamined search systems to the comparison set. Until now, studies have largely examined the suitability of search systems for certain scholarly tasks for individual systems or by comparing a few systems.[33,50,95] Our methods allowed a comprehensive review of many different search systems. As a result we found significant performance differences among the search engines examined, confirming that no single search system is perfect. Their efficient use thus demands searchers are well-trained and can weigh up a system's strengths and weaknesses and make informed decisions on where and how to search. Only then can reviewers evaluate search systems and match our tested search system suggestions to their subjective information requirements.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Appendix (see Supporting Information).

### ACKNOWLEDGEMENTS

We like to thank the two anonymous reviewers, associate editor, and editor Prof Gerta Rücker of Research Synthesis Methods for their valuable comments that helped improving this paper. All remaining errors and omissions are ours.

### CONFLICT OF INTEREST

The author reported no conflict of interest.

### ORCID

*Michael Gusenbauer* https://orcid.org/0000-0001-7768-2351

*Neal R. Haddaway* https://orcid.org/0000-0003-3902-2234

### REFERENCES

1. Price DJS. *Little Science, Big Science*. New York: Columbia Univ. Press; 1963 Columbia paperback.
2. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation

Index. *Scientometrics.* 2010;84(3):575-603. https://doi.org/10.1007/s11192-010-0202-z
3. Eden D. From the editors: replication, meta-analysis, scientific progress, and AMJ's publication policy. *AMJ.* 2002;45(5):841-846. https://doi.org/10.5465/AMJ.2002.7718946
4. Naisbitt J, Aburdene P. *Megatrends 2000: Ten New Directions for the 1990's*. 1st ed. New York: Morrow; 1990.
5. Cooper HM. *Research Synthesis and Meta-analysis: A Step-by-Step Approach*. Applied social research methods series. Fifth ed. 2 Los Angeles: SAGE; 2017.
6. Littell JH. *Conceptual and Practical Classification of Research Reviews and Other Evidence Synthesis Products*; 2018.
7. Kostoff RN, Shlesinger MF. CAB: citation-assisted background. *Scientometrics.* 2005;62(2):199-212. https://doi.org/10.1007/s11192-005-0014-8
8. Littell JH, Corcoran J, Pillai V. *Systematic Reviews and Meta-Analysis:* Oxford University Press, USA; 2008. https://books.google.at/books?id=UpsRDAAAQBAJ.
9. Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, Brigham TJ. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. *J Clin Epidemiol.* 2015;68(6):617-626. https://doi.org/10.1016/j.jclinepi.2014.11.025
10. Bandara W, Furtmueller E, Gorba, Gorbacheva E, Miskon S, Beekhuyzen J. Achieving rigor in literature reviews: insights from qualitative data analysis and tool-support. *Communications of the Association for Information Systems.* 2015;37(8):154-204. http://aisel.aisnet.org/cais/vol37/iss1/8
11. Meert D, Torabi N, Costella J. Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *J Med Libr Assoc.* 2016;104(4):267-277. https://doi.org/10.3163/1536-5050.104.4.004
12. Koffel JB. Use of recommended search strategies in systematic reviews and the impact of librarian involvement: a cross-sectional survey of recent authors. *Plos One.* 2015;10(5):1-13. https://doi.org/10.1371/journal.pone.0125931
13. Livoreil B, Glanville J, Haddaway NR, et al. Systematic searching for environmental evidence using multiple tools and sources. *Environ Evid.* 2017;6:1-14. https://doi.org/10.1186/s13750-017-0099-6
14. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*; 2011; Version 5.1.0.
15. Kugley S, Wade A, Thomas J, et al. *Searching for studies: GUidelines on information retrieval for Campbell Systematic Reviews*; 2016; 1.
16. Pullin A, Frampton G, Livoreil B, Petrokofsky G. *Guidelines and Standards for Evidence Synthesis in Environmental Management. Version 5.0*; 2018.
17. Hug SE, Braendle MP. The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics.* 2017;113(3):1551-1571. https://doi.org/10.1007/s11192-017-2535-3
18. Khabsa M, Giles CL. The number of scholarly documents on the public web. *Plos One.* 2014;9(5):1-6. https://doi.org/10.1371/journal.pone.0093949
19. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB journal.* 2008;22(2):338-342. https://doi.org/10.1096/fj.07-9492LSF
20. Orduña-Malea E, Ayllón JM, Martín-Martín A, Delgado L-CE. Methods for estimating the size of Google Scholar. *Scientometrics.* 2015;104(3):931-949. https://doi.org/10.1007/s11192-015-1614-6
21. Harzing A-W. A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics.* 2014;98(1):565-575. https://doi.org/10.1007/s11192-013-0975-y
22. Meier JJ, Conkling TW. Google Scholar's coverage of the engineering literature: an empirical study. *The Journal of Academic Librarianship.* 2008;34(3):196-201. https://doi.org/10.1016/j.acalib.2008.03.002
23. Turnbull D, Berryman J. *Relevant search: with applications for Solr and Elasticsearch*. Shelter Island New York: Manning Publications Co; 2016.
24. Levay P, Ainsworth N, Kettle R, Morgan A. Identifying evidence for public health guidance: a comparison of citation searching with Web of Science and Google Scholar. *Res Synth Methods.* 2016;7(1):34-45. https://doi.org/10.1002/jrsm.1158
25. Boeker M, Vach W, Motschall E. Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Med Res Methodol.* 2013;13:1-12. https://doi.org/10.1186/1471-2288-13-131
26. Hjørland B. Classical databases and knowledge organization: a case for Boolean retrieval and human decision-making during searches. *J Assn Inf Sci Tec.* 2015;66(8):1559-1575. https://doi.org/10.1002/asi.23250
27. Weber K. Search engine bias. In: Lewandowski D, ed. *Handbuch Internet-Suchmaschinen 2*. AKA Verlag Heidelberg; 2011:265-285.
28. Vaughan L, Thelwall M. Search engine coverage bias: Evidence and possible causes. *Information Processing & Management.* 2004;40(4):693-707. https://doi.org/10.1016/S0306-4573(03)00063-3
29. Carmines EG, Zeller RA. *Reliability and Validity Assessment*. Quantitative applications in the social sciences. Beverly Hills, London: Sage Publications; 1979 no.07–017.
30. Jacsó P. Google Scholar: the pros and the cons. *Online Information Review.* 2005;29(2):208-214. https://doi.org/10.1108/14684520510598066
31. Jacsó P. Google Scholar revisited. *Online Information Review.* 2008;32(1):102-114. https://doi.org/10.1108/14684520810866010
32. Bethel A, Rogers M. A checklist to assess database-hosting platforms for designing and running searches for systematic reviews. *Health Info Libr J.* 2014;31(1):43-53. https://doi.org/10.1111/hir.12054
33. Bramer WM, Giustini D, Kramer B, Anderson P. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Syst Rev.* 2013;2:1-9. https://doi.org/10.1186/2046-4053-2-115
34. Bramer WM, Giustini D, Kramer BMR. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. *Syst Rev.* 2016;5:1-9. https://doi.org/10.1186/s13643-016-0215-7
35. Mowshowitz A, Kawaguchi A. Measuring search engine bias. *Information Processing & Management.* 2005;41(5):1193-1205. https://doi.org/10.1016/j.ipm.2004.05.005

36. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6(7):1-6. https://doi.org/10.1371/journal.pmed.1000097
37. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics.* 2019;118(1):177-214. https://doi.org/10.1007/s11192-018-2958-5
38. Ortega JL. *Academic Search Engines: A quantitative outlook*. Chandos information professional series. Oxford, UK: Chandos Publishing/Elsevier; 2014.
39. Schöpfel J, Farace DJ. Grey literature. In: Bates MJ, Maack MN, eds. *Encyclopedia of library and information sciences*. 3rd ed. / edited by Boca Raton, Fla: CRC London: Taylor & Francis; 2010:2029–2039.Marcia J. Bates and Mary Niles Maack
40. Sampson M, McGowan J. Inquisitio validus Index Medicus: a simple method of validating MEDLINE systematic review searches. *Res Synth Methods.* 2011;2(2):103-109. https://doi.org/10.1002/jrsm.40
41. Rogers M, Bethel A, Abbott R. Locating qualitative studies in dementia on MEDLINE, EMBASE, CINAHL, and PsycINFO: a comparison of search strategies. *Res Synth Methods.* 2017;9(2):579-586. https://doi.org/10.1002/jrsm.1280
42. Rader T, Mann M, Stansfield C, Cooper C, Sampson M. Methods for documenting systematic review searches: a discussion of common issues. *Res Synth Methods.* 2014;5(2):98-115. https://doi.org/10.1002/jrsm.1097
43. O'Mara-Eves A, Brunton G, McDaid D, Kavanagh J, Oliver S, Thomas J. Techniques for identifying cross-disciplinary and 'hard-to-detect' evidence for systematic review. *Res Synth Methods.* 2014;5(1):50-59. https://doi.org/10.1002/jrsm.1094
44. Atkinson KM, Koenka AC, Sanchez CE, Moshontz H, Cooper H. Reporting standards for literature searches and report inclusion criteria: making research syntheses more transparent and easy to replicate. *Res Synth Methods.* 2015;6(1):87-95. https://doi.org/10.1002/jrsm.1127
45. Mahood Q, van Eerd D, Irvin E. Searching for grey literature for systematic reviews: challenges and benefits. *Res Synth Methods.* 2014;5(3):221-234. https://doi.org/10.1002/jrsm.1106
46. Bar-Ilan J. On the overlap, the precision and estimated recall of search engines. A case study of the query "Erdos". *Scientometrics.* 1998;42(2):207-228. https://doi.org/10.1007/BF02458356
47. Kumar BTS, Prakash JN. Precision and relative recall of search engines: a comparative study of Google and Yahoo. *Singapore Journal of Library and Information Management*. 2009;38:124-137.
48. Shafi SM, Rather R. Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology.* 2005;2(2):1-7.
49. Usmani TA, Pant D, Bhatt AK. A comparative study of Google and Bing search engines in context of precision and relative recall parameter. *International Journal on Computer Science and Engineering (IJCSE).* 2012;4(1):21-34.
50. Giustini D, Boulos MNK. Google Scholar is not enough to be used alone for systematic reviews. *Online J Public Health Inform.* 2013;5(2):1-10. https://doi.org/10.5210/ojphi.v5i2.4623
51. Bramer WM. Variation in number of hits for complex searches in Google Scholar. *Journal of the Medical Library Association.* 2016;104(2):143-145. https://doi.org/10.3163/1536-5050.104.2.009
52. Brophy J, Bawden D. Is Google enough? Comparison of an internet search engine with academic library resources. *Aslib Proceedings.* 2005;57(6):498-512. https://doi.org/10.1108/00012530510634235
53. Gehanno J-F, Rollin L, Darmoni S. Is the coverage of Google Scholar enough to be used alone for systematic reviews. *BMC Medical Informatics and Decision Making.* 2013;13(7):1-5.
54. Sturm B, Sunyaev A. If you want your research done right, do you have to do it all yourself? Developing design principles for systematic literature search systems. In: Maedche A, Vom Brocke J, Hevner A, eds. *Designing the Digital Transformation*; 2017:138–146.
55. Chu H, Rosenthal M. Search engines for the World Wide Web: a comparative study and evaluation methodology. *J. Am. Soc. Inf. Sci.* 1996;33:127-135.
56. Biolcati-Rinaldi F, Molteni F, Salini S. Assessing the reliability and validity of Google Scholar indicators. The case of social sciences in Italy. In: Bonaccorsi A, ed. *The Evaluation of Research in Social Sciences and Humanities: Lessons from the Italian Experience*. Cham: Springer International Publishing; 2018:295-319.
57. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *Plos One.* 2015;10(9):1-17. https://doi.org/10.1371/journal.pone.0138237
58. Aune D, Giovannucci E, Boffetta P, et al. Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose–response meta-analysis of prospective studies. *International Journal of Epidemiology.* 2017;46(3):1029-1056. https://doi.org/10.1093/ije/dyw319
59. Barnett DW, Barnett A, Nathan A, van Cauwenberg J, Cerin E. Built environmental correlates of older adults' total physical activity and walking: a systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity.* 2017;14(1):1-24. https://doi.org/10.1186/s12966-017-0558-z
60. Baur D, Gladstone BP, Burkert F, et al. Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and *Clostridium difficile* infection: a systematic review and meta-analysis. *Lancet Infectious Diseases.* 2017;17(9):990-1001. https://doi.org/10.1016/S1473-3099(17)30325-0
61. Bediou B, Adams DM, Mayer RE, Tipton E, Green CS, Bavelier D. Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin.* 2018;144(1):77-110. https://doi.org/10.1037/bul0000130
62. Bethel MA, Patel RA, Merrill P, et al. Cardiovascular outcomes with glucagon-like peptide-1 receptor agonists in patients with type 2 diabetes: a meta-analysis. *Lancet Diabetes & Endocrinology.* 2018;6(2):105-113. https://doi.org/10.1016/S2213-8587(17)30412-6
63. Bourne RRA, Flaxman SR, Braithwaite T, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a

systematic review and meta-analysis. *Lancet Global Health.* 2017;5(9):E888-E897. https://doi.org/10.1016/S2214-109X(17)30293-0

64. Brunoni AR, Chaimani A, Moffa AH, et al. Repetitive transcranial magnetic stimulation for the acute treatment of major depressive episodes a systematic review with network meta-analysis. *Jama Psychiatry.* 2017;74(2):143-152. https://doi.org/10.1001/jamapsychiatry.2016.3644
65. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlof E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy.* 2018;47(1):1-18. https://doi.org/10.1080/16506073.2017.1401115
66. Chu DK, Kim LH-Y, Young PJ, et al. Mortality and morbidity in acutely ill adults treated with liberal versus conservative oxygen therapy (IOTA): a systematic review and meta-analysis. *Lancet.* 2018;391(10131):1693-1705. https://doi.org/10.1016/S0140-6736(18)30479-3
67. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet.* 2018;391(10128):1357-1366. https://doi.org/10.1016/S0140-6736(17)32802-7
68. Stacey D, Légaré F, Lewis K, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev.* 2017;4:1-343. https://doi.org/10.1002/14651858.CD001431.pub5
69. *Oxford Wordlist*. Oxford University Press; 2008.
70. Fieschi M, Coiera E, Li YCJ. *Medinfo*: IOS Press; 2004. https://books.google.at/books?id=bS2xdt7iufgC.
71. Meho LI, Yang K. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *J. Am. Soc. Inf. Sci.* 2007;58(13):2105-2125. https://doi.org/10.1002/asi.20677
72. Martín-Martín A, Orduna-Malea E, Thelwall M, López-Cózar ED. Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. *Journal of Informetrics.* 2018;12(4):1160-1177. https://doi.org/10.31235/osf.io/42nkm
73. Bakkalbasi N, Bauer K, Glover J, Wang L. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomed Digit Libr.* 2006;3(7):1-8. https://doi.org/10.1186/1742-5581-3-7
74. Haddaway NR, Macura B, Whaley P, Pullin AS. ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environ Evid.* 2018;7(7):1-8. https://doi.org/10.1186/s13750-018-0121-7
75. White RW, Roth RA. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool; 2009.
76. Jansen BJ, Spink A. In: Langendoerfer P, Droegehorn O, eds. IC'2003*An Analysis of Web Documents Retrieved and Viewed*; 2003:65-69.
77. Jansen BJ, Spink A. How are we searching the World Wide Web?: a comparison of nine search engine transaction logs. *Information Processing & Management.* 2006;42(1):248-263. https://doi.org/10.1016/j.ipm.2004.10.007
78. Athukorala K, Głowacka D, Jacucci G, Oulasvirta A, Vreeken J. Is exploratory search different?: a comparison of information search behavior for exploratory and lookup tasks. *J Assn Inf Sci Tec.* 2016;67(11):2635-2651. https://doi.org/10.1002/asi.23617
79. Hemminger BM, Lu D, Vaughan KTL, Adams SJ. Information seeking behavior of academic scientists. *J. Am. Soc. Inf. Sci.* 2007;58(14):2205-2225. https://doi.org/10.1002/asi.20686
80. Athukorala K, Hoggan E, Lehtiö A, Ruotsalo T, Jacucci G. Information-seeking behaviors of computer scientists: challenges for electronic literature search tools. *Proc. Am. Soc. Info. Sci. Tech.* 2013;50(1):1-11. https://doi.org/10.1002/meet.14505001041
81. Nicholas D, Boukacem-Zeghmouri C, Rodríguez-Bravo B, et al. Where and how early career researchers find scholarly information. *Learned Publishing.* 2017;30(1):19-29. https://doi.org/10.1002/leap.1087
82. Niu X, Hemminger BM. A study of factors that affect the information-seeking behavior of academic scientists. *J. Am. Soc. Inf. Sci.* 2012;63(2):336-353. https://doi.org/10.1002/asi.21669
83. Sapa R, Krakowska M, Janiak M. Information seeking behaviour of mathematicians: scientists and students. *Information Research: An International Electronic Journal*. 2014;19(4):1-11.
84. Fast KV, Campbell DG. "I still like Google": University student perceptions of searching OPACs and the web. *Proceedings of the American Society for Information Science and Technology.* 2004;41(1):138-146. https://doi.org/10.1002/meet.1450410116
85. Kuiper E, Volman M, Terwel J. Students' use of Web literacy skills and strategies: searching, reading and evaluating Web information. *Information Research.* 2008;13(3):1-18.
86. Ingwersen P. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation.* 1996;52(1):3-50. https://doi.org/10.1108/eb026960
87. Wellings S, Casselden B. An exploration into the information-seeking behaviours of engineers and scientists. *Journal of Librarianship and Information Science.* 2017;9(2):1-12. https://doi.org/10.1177/0961000617742466
88. Rowlands I, Nicholas D, Williams P, et al. The Google generation: the information behaviour of the researcher of the future. *AP.* 2008;60(4):290-310. https://doi.org/10.1108/00012530810887953
89. Kingsley K, Galbraith GM, Herring M, Stowers E, Stewart T, Kingsley KV. Why not just Google it? An assessment of information literacy skills in a biomedical science curriculum. *BMC Med Educ.* 2011;11(17):1-8. https://doi.org/10.1186/1472-6920-11-17
90. Kurbanoğlu S, Boustany J, Špiranec S, Grassian E, Mizrachi D, Roy L, eds. *Information literacy: moving toward sustainability: Third European conference, ECIL 2015, Tallinn, Estonia, October 19–22, 2015: revised selected papers.* Cham, Heidelberg, New York: Springer; 2015. Communications in computer and information science; 552.
91. Kurbanoğlu S, Boustany J, Špiranec S, et al. (Eds). *Search Engine Literacy: Information Literacy in the*. Workplace: Springer International Publishing; 2018.
92. Brindesi H, Monopoli M, Kapidakis S. Information seeking and searching habits of Greek physicists and astronomers: a case study of undergraduate students. *Procedia - Social and*

*Behavioral Sciences*. 2013;73:785-793. https://doi.org/10.1016/j.sbspro.2013.02.119

93. Kirschner PA, de Bruyckere P. The myths of the digital native and the multitasker. *Teaching and Teacher Education*. 2017;67:135-142. https://doi.org/10.1016/j.tate.2017.06.001
94. Georgas H. Google vs. the library (part II): student search patterns and behaviors when using Google and a federated search tool. *Portal: Libraries and the Academy*. 2014;14(4):503-532.
95. Halevi G, Moed H, Bar-Ilan J. 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*. 2017;11(3):823-834. https://doi.org/10.1016/j.joi.2017.06.005

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res Syn Meth*. 2020;11:181–217. https://doi.org/10.1002/jrsm.1378

International Entrepreneurship and Management Journal (2020) 16:1023–1042
https://doi.org/10.1007/s11365-020-00635-4

# The art of crafting a systematic literature review in entrepreneurship research

Sascha Kraus[1] · Matthias Breier[2] · Sonia Dasí-Rodríguez[3]

Published online: 1 February 2020
© The Author(s) 2020

## Abstract

Systematic literature reviews are an increasingly used review methodology to synthesize the existing body of literature in a field. However, editors complain about a high number of desk rejections because of a lack in quality. Poorly developed review articles are not published because of a perceived lack of contribution to the field. Our article supports authors of standalone papers and graduate students in the Entrepreneurship domain to write contribution-focused systematic reviews e.g. by providing a concrete guideline. Our article analyzes the strengths and weaknesses of a systematic literature review and how they can be overcome. Furthermore, we provide a combined list of highly ranked journals in the Entrepreneurship domain as a basis for quality appraisal. Finally, this article builds a scenario for the future of the systematic literature review methodology and shows how technological improvements have changed this methodology and what can be achieved in the future.

**Keywords** Systematic literature review · Structured literature review · Entrepreneurship · Journal rankings · State-of-the-art

✉ Sascha Kraus
sascha.kraus@durham.ac.uk

Matthias Breier
matthias.breier@gmx.at

Sonia Dasí-Rodríguez
sonia.dasi@uv.es

[1] Durham University, Durham University Business School, Mill Hill Lane, Durham DH1 3LB, UK

[2] Lappeenranta University of Technology, School of Business, Skinnarilankatu 34, 53850 Lappeenranta, Finnland

[3] Universitat de València, València, Spain

## Introduction

The *systematic (also known as "structured") literature review* (SLR) entered management research as a promising methodology for reviewing previous literature to bring the field closer together (Tranfield et al. 2003). High quality SLRs support better decisions for policy-makers and entrepreneurs and help researchers to synthesize the literature under review. Due to the rapidly growing popularity of this methodology within the overall management domain, the SLR received more attention, and initial papers created rules and suggestions on how to conduct such review articles in the field. The SLR became increasingly popular, almost even replacing traditional reviews for individual review papers in management (Jones and Gatrell 2014). The main advantages of an SLR are transparency in data collection and synthesis that results in a higher level of objectivity and reproducibility (Tranfield et al. 2003).

When talking about *traditional* literature reviews, we basically mean non-structured, −systematic or -transparent reviews with a higher level of subjectivity in data-collection and data-interpretation. Systematic reviews are also not novel. The first examples emerged at the end of the nineteenth century when the term *systematic review* was used for nearly every review article (Petticrew and Roberts 2006). These reviews changed a lot over time. If, for example, we compare the situation Hart (1998) describes with today, we see major differences. This is most evident in the increased availability of literature in general, especially online databases (such as EBSCO, Scopus, ABI/Inform etc.) allowing researchers to conduct quicker and much more transparent review processes than was possible in the final years of the last century. However, while Hart mainly regards the SLR as one method of giving an overview on the topic e.g. within a larger research project, they have also established themselves as a standalone methodology by itself to create evidence in a topic. In 1998, Hart criticized a lot of the review articles published at that time for their lack in quality, and we often face the same problem today; for example, in traditional *Research Methods* courses at universities, the SLR does not receive the same attention as statistical methods, meaning that a significant cohort of the research community is more dedicated to subjectively summarizing the extant literature in their field than synthesizing it in a systematic manner.

With this article, we focus on the SLR in Entrepreneurship research – more specifically its delimitations, the state-of-the-art, its limitations and outlook for this methodology. We also provide advice which is best practice. Our article addresses researchers/scholars, Ph.D. candidates and graduate students – especially in the Entrepreneurship domain, as it specifically includes the field specifications in the discussion. Our aim is to discuss potential further developments based on the SLR in order to overcome the main limitations of the methodology by comparing it with other relevant review methods and setting it in the context of the Entrepreneurship domain.

With this contribution to Entrepreneurship research, we aim to further support the importance of the synthesis of previous works and help to consolidate the existing body of literature in a certain field. SLRs offer the possibility of combining existing literature and create solid definitions and foundations for further research. This study supports the strengthening of SLRs as a methodology in Entrepreneurship. In science, there are several reasons to write a literature review. One of these is for the literature review to set a beginning for a dissertation and detect research gaps and questions that can be answered by the dissertation. Last, but not least, our article aims to provide (particularly

young) scholars in Entrepreneurship (and related areas) at all academic levels with a specific guideline to carry out their literature reviews in Entrepreneurship.

## The need for reviewing the literature

In research, there are different aims and situations in which scholars write a literature review. Typically, the aim of literature reviews is to summarize and integrate the existing knowledge about a topic (Rowley and Slack 2004). The situations where literature reviews are written and used are independent from the authors that deal with the topic. In broad terms, there are three main situations for writing a literature review (Knopf 2006; Okoli 2015):

1. A standalone review article of the literature for a specific topic.
2. An introduction to an empirical paper and foundation for hypotheses.
3. The first stage of a bigger research project (e.g. a dissertation).

For this article, we mainly concentrate on the first of these, the *standalone SLR* (Okoli 2015). However, for a dissertation (especially at doctoral level), we believe that an SLR is more useful than a traditional literature review. SLRs are a specific methodology that allow for the creation of a whole article based on reviewing the literature without collecting empirical data. They try to answer a research question, usually about the status quo of a field of research. In empirical articles, on the other hand, a literature review does not need to answer a research question on its own; instead, it provides a short overview of the topic and helps to derive the main hypotheses of the paper. Therefore, a traditional literature review might in some cases be more suitable for empirical articles, where the main focus is not the literature review itself.

We can differentiate two different streams of literature reviews: traditional (subjective) and structured reviews (see Table 1). Unlike SLRs, traditional literature reviews do not follow a reproducible and transparent methodology.

### The traditional literature review

The majority of literature reviews in management research are not systematic by nature, i.e. they follow a narrative rather than a transparent methodology (Briner and Denyer 2012). They are thus often regarded to be "unscientific" because of their use of unrepresentative samples and unsystematic procedures (Mulrow 1994; Oakley 2002). Traditional literature reviews often do not evaluate the quality of articles nor do they follow any specific rules. They are often used in introductions to empirical papers or conference presentations to support the hypotheses the author has developed. Therefore, the authors are often less interested in showing previous studies which are contradictory to their intended hypotheses, and create a bias by omitting them. In "cumulative" (article-based) Ph.D. dissertations, traditional reviews are usually used to tie the topic together and create a "bracker" around the individual papers.

Because of the lack of systematization and transparency, traditional reviews are more likely to be biased by the subjectivity of the author (Hodgkinson and Ford 2014; Mulrow 1994). The difference between traditional and systematic reviews mostly lies

in the methods of data collection and possibilities of replication. Traditional reviews do not follow a strict rule of how studies are collected and therefore it is highly possible that the study selection is driven by the subjectivity of the author. The same issue is criticized when it comes to quality control of the studies. In both cases, the driving factor that takes the final decision is the authors themselves (Tranfield et al. 2003).

Nevertheless, traditional literature reviews have several advantages. One of the most important reasons for writing one is that it is written more easily than an SLR. Traditional reviews are driven by the intuition and the experience of the authors, but they are also subjectively influenced by them. Thanks to advances in electronic databases, SLRs today can be conducted quicker and more transparently than was previously possible. Previously, even if researchers created a transparent process, their ability to search for keywords to uncover contributions from other fields was limited. Thus, subjective literature reviews were important early on and therefore have a long tradition in science. Hart (1998) gives a useful overview on literature reviews. In her article, the author outlines what constitutes a good literature review, which from today's perspective, differs significantly from an SLR. However, this overview does not separate subjective reviews from SLRs.

## The systematic literature review

### Origin and development of systematic literature reviews

An SLR is a form of research that deals with existing publications and follows a systematic methodology for synthesizing data that is already published (Tranfield et al. 2003). An SLR follows a pre-defined process to analyze literature in a reproducible manner. Besides its transparent methodology, SLRs rank literature by its quality. SLRs tend to follow a research question and aim to answer it in the best way. They reach a conclusion and exhibit knowledge about a specific topic in research. SLRs have their own limitations and contrary to traditional literature reviews, they admit to these limitations and transparently list them (Frank and Hatak 2014). As the transparent process allows a reproducible methodology and all necessary literature is integrated, this is a sound basis to draw conclusions and create evidence.

Based on the descriptions above we define a SLR as follows:

> *"An SLR is a review of an existing body of literature that follows a transparent and reproducible methodology in searching, assessing its quality and synthesizing it, with a high level of objectivity."*

Systematic reviews are not new per se, however they have changed dramatically in the recent years. At the end of the nineteenth century, the term "systematic review" was used for almost every review article. With the rise of *meta-analysis* as a more advanced (statistical) review methodology, the field was elevated to a new standard. Policy-makers identified the potential of systematic reviews for evidence-based decision making (Petticrew and Roberts 2006). The evidence-based approach has been adapted from the field of medicine, which suffered from a major increase in the body of literature so that existing studies e.g. on one treatment had to be aligned to derive evidence, and has generally led to an increase in the value of systematic reviews (Ohlsson 1994). Based on a broad investigation of the similarities and differences of

medicine and management, Tranfield et al. (2003) developed a first overview to support the rise of systematic reviews in management research. The publication of this article can be regarded as the tipping point for SLR in management.

With regard to evidence-based decision-making, systematic reviews are considered to be a powerful form of research. Tranfield et al. (2006) refer to a table of the hierarchy of evidence in health care which shows that systematic reviews are considered the highest standard of evidence. Individual studies can show different results for the same or similar issues; however, a systematic review provides an overview of all these individual studies. Petticrew and Roberts (2006) compare an individual study with an individual answer in a quantitative study. Only the sum of more than one answer can help to overcome biases and create a reliable answer on the hypotheses. For SLRs, this is the same, resulting in a high level of confidence in the review article when it comes to answering questions or hypotheses.

There are also differences among SLRs themselves. A SLR can significantly vary in quality which is dependent on the authors. Later in this article, we analyze common mistakes an author can make in his SLR. Further differences occur in the chosen topic. Based on the amount of literature which is available to conduct a review article, the author must find a balance between width and depth. Another important factor is the heterogeneity an author allows within the search process (Frank and Hatak 2014).

## First steps for a systematic literature review

Before beginning an SLR, the author always has to ask if and why a SLR is needed or if a traditional review is adequate (Pittaway et al. 2014). If the author sees a

**Table 1** Comparison between systematic and traditional review

| | Systematic Literature Review | Traditional Literature Review |
|---|---|---|
| Identification for the need for a review article | SLRs only make sense if there is a need for one. | Traditional reviews are part of nearly every publication. |
| Development of a review protocol | Essential for the objectivity of an SLR | Not common for traditional reviews |
| Identification of research | Structured, replicable and transparent process | Subjective process |
| Evaluating studies | Transparent protocol of eliminated studies, objective process of elimination | The author takes the literature that helps to support their hypotheses |
| Conducting data extraction | Driven by a general protocol | Driven by the intuition of the authors |
| Conducting data synthesis | Concept driven; central part of an SLR | Not necessary; resembles a summary of existing literature |
| Reasons for a SLR | Standalone paper, creates evidence and answer a research question | To set a literature foundation for an upcoming empirical project, as individual paper (formerly) |
| Labor costs /Time | Very time consuming | Less consuming than an SLR |

requirements for an SLR, a research question needs to be specified (Briner and Denyer 2012). The research question is one of the general differences among traditional literature reviews and SLRs. The evidence the SLR aims to create is highly connected to the research question specified at the very beginning. Oftentimes, the main research question of a SLR is to synthesize what we know and what we do not know about a research question, hypotheses, applied methods or topics (Briner and Denyer 2012), i.e. to offer a state-of-the-art overview of the research in a current field, which identifies research gaps and potentially even already develops an own research model which might then be used as a basis for further (empirical) research. As a "good practice examples", we would recommend e.g. the SLR about entrepreneurial intentions by Liñán and Fayolle (2015) or the one about innovation in family firms by Calabrò et al. (2019).

The basis to write a SLR is the availability of sufficient literature on the subject to justify a synthesis (Hodgkinson and Ford 2015). A literature review on a very narrow niche with only a very limited amount of papers can only very rarely provide new insights or theories. Alternatively, in research areas with a broad range of literature which is fragmented and based on inconsistent terminologies, a systematic review can help to consolidate the topic in the sense of a "status quo of current research". In other words: When a research field is rather new in itself, and mostly case studies or qualitative research is present, and only very few quantitative research, then all identified sources should be analyzed – whereas when there is already a sufficient amount of quantitative research being published (i.e. when the research field in itself has already evolved from theory development to theory testing), previous qualitative research might be excluded from the literature review.

Another reason to conduct an SLR is the lack of a rugged theoretical design, as a good and carefully developed SLR have the ability to create new insights in form of a new theoretical construct (Pittaway et al. 2014).

By contrast, an SLR should not be conducted if it is not the right methodology to answer the research question. In research areas where good SLRs are published recently, there might also be a lack of further knowledge that can be synthesized. Exceptionally, an SLR can be a useful tool if the existing review lacks a good synthesis or if groundbreaking progress has been made in the interim period. SLRs should also be avoided if the research question that should be answered is not detailed enough or if it sets the wrong focus (Petticrew and Roberts 2006).

### Contributions of systematic literature reviews

In Entrepreneurship, as a still comparatively young discipline (Ferreira et al. 2019), sub-research fields developed quickly in the last decades (Pittaway et al. 2014). Therefore, many papers are published from authors with different backgrounds using slightly different terminologies for the same research object or same terminologies for different objects. Therefore, within a certain research field, authors with diverse backgrounds see constructs, theories and so on with their own eyes, which results in a scattered field. SLRs can help to overcome this issue by synthesizing the field and creating a common language. On this foundation, an SLR can increase awareness within the field and show current perspectives (Frank and Hatak 2014) even from different disciplines and backgrounds (Pittaway et al. 2014).

When conducting an SLR, the author gains an overview of the most important literature on the topic. On this basis, the author has the opportunity to create a map of all the knowledge that was gathered in the field (Armitage and Keeble-Allen 2008; Frank and Hatak 2014). As the process of collecting literature is standardized by the author on basis of hit keywords, he can also discover publications from other disciplines. This knowledge map and the broad view from different disciplines help to give a holistic view and provide the foundation to synthesize the research field across the disciplines (Pittaway et al. 2014). The creation of a knowledge map can also help the author and other researchers to set a personal research focus and define a niche that enables the development of new research (Tranfield et al. 2003). Furthermore, the knowledge map can show growing research trends and directions of the field.

Besides the positive effects of the knowledge map outlined above, it can help to further develop a theory. A broad overview on a specific topic with the influence of different disciplines can help researchers to devise new theories. This result can be one of the main aims of an SLR (Pittaway et al. 2014). The creation of new theoretical constructs may lead to new directions in the research area and support the overall discussion (Frank and Hatak 2014).

The period since the late 1990s has seen more and more evidence-based movements arise. Their objective is to take decisions based on evidence from science. SLRs and meta-analyses are the primary evidence producers in other disciplines too. Systematic reviews are largely seen as strong evidence (Table 2), as they are less vulnerable to error and bias (Tranfield et al. 2003). Briner and Denyer (2012) summarize that systematic reviews are important as they create evidence to support decisions in research and practice. This linkage from research to practice is a reason why SLRs are used for research funding requests (Frank and Hatak 2014).

A survey among scientists in America showed that 0.3% falsify their data (Martinson et al. 2005). Therefore, the process of systematically reviewing a bulk of literature constitutes an important quality test for published articles. While peer reviews basically check the article, but not the data itself, they do not have the ability to check for manipulation. A systematic review might thus help to uncover manipulation or at least criticize individual studies. So, if the author deals with several studies that all show the same results but there is a study that shows significantly different results, suspicion may be raised and the study can be investigated in more detail (Rousseau et al. 2008). The author should check how the authors try to justify their results.

SLRs and (statistical) meta-analyses are generally supposed to constitute the highest possible evidence. However, within Entrepreneurship as a social science, the

**Table 2** Hierarchy of evidence in medicine (Davies and Nutley 1999)

| | |
|---|---|
| I-I | Systematic review and meta-analysis of two or more double blind randomized controlled trials, |
| I-2 | One or more large double-blind randomized controlled trials, |
| II-1 | One or more well-conducted cohort studies, |
| II-2 | One or more well-conducted case-control studies, |
| II-3 | A dramatic uncontrolled experiment, |
| III-1 | Expert committee sitting in review; peer leader opinion, |
| IV | Personal experience. |

elimination of a bias is not expected to be possible and therefore accepted. So no final evidence is possible (Denyer and Tranfield 2006). The authors themselves remain a limitation. Although, the transparent process tries to minimize the subjective influence, the authors' bias remains. Basically, authors are never completely objective when they review the literature.

The SLR methodology has received increased attention in Entrepreneurship research in recent years. Database searches on the keywords "Entrepreneurship" and "Systematic Review" typically yield hundreds of publications. Hence, there is an extensive and growing list of high quality SLRs in Entrepreneurship. The best practice is to check review articles published in journals with a high ranking (see Table 3) and journals that specialized on review articles.

## About the entrepreneurship research domain

Research on Entrepreneurship is still getting more and popular today, and the domain has grown significantly in recent years. The sophistication of the research domain established new sub-fields. Family business management, entrepreneurial behavior, small business management, female entrepreneurship, technology entrepreneurship or social entrepreneurship – just to mention a few – emerged as their own sub-fields. In the coming chapter, we provide a broad overview of the specific top journal outlets of this Entrepreneurship research domains, and list all Entrepreneurship journals ranked in the three major academic journal rankings *VHB Jourqual (JQ) 3* (Germany) from 2015, the *Academic Journal Guide/ABS* (UK) from 2018 and the *JCR Impact Factors* (IF) by Clarivate Analytics from 2018.

Some journals are ranked in different sub-categories within the rankings. For example, the *Journal of Technology Transfer* is part of the "Entrepreneurship" section in JQ3, while it is part of the section "Innovation" in the ABS. If a journal is mentioned in the section of Entrepreneurship in at least one of the rankings, it is inserted into the table.

The three journal rankings follow different methodologies. Bouncken et al. (2015) created a transformation table to compare the rankings (Table 4), which we have adapted accordingly to account for the increase in the Impact Factors over the recent years. The quality criteria of the journals can be used for quality appraisal, as well as for limiting down the number of publications if the search strings result in too many publications by the application of a "quality threshold" (e.g. of only including a publication in a target journal which is rated "C" or equivalent in at least one of the three rankings).

In the last few years, more and more journals started to specifically publish review articles. For example, the *International Journal of Management Reviews*, the *Academy of Management Review* and *Management Review Quarterly* specifically focused on review articles, whereas some other journals (such as the e.g. the *Review of Managerial Science*) have their own categories for review articles (as opposed to "regular articles" – in the submission system). Some journals, such as the *Journal of Management* or the *Journal of Business Research*, have published distinct review issues which deal with literature review articles only. SLRs in general and journals focused on review articles in particular, usually generate above average citation counts with these issues.

**Table 3** Top Entrepreneurship Journals based on VHB, ABS and JCR (as of 2020)

| No. | Journal | VHB JQ3 | ABS | JCR IF |
|---|---|---|---|---|
| 1 | Journal of Business Venturing | A | 4 | 6.333 |
| 2 | Entrepreneurship: Theory and Practice | A | 4 | 6.193 |
| 3 | Strategic Entrepreneurship Journal | A | 4 | 2.956 |
| 4 | Family Business Review | B | 3 | 6.188 |
| 5 | Small Business Economics | B | 3 | 3.555 |
| 6 | Journal of Small Business Management | B | 3 | 3.120 |
| 7 | Entrepreneurship & Regional Development | B | 3 | 2.928 |
| 8 | Journal of Technology Transfer | B | 2 | 4.037 |
| 9 | Industry & Innovation | B | 2 | 3.157 |
| 10 | International Journal of Innovation Management | B | 2 | – |
| 11 | International Journal of Entrepreneurial Venturing | B | 1 | – |
| 12 | International Entrepreneurship and Management Journal | B/C | 1 | 2.537 |
| 13 | Technovation | C | 3 | 5.250 |
| 14 | International Small Business Journal | C | 3 | 3.706 |
| 15 | International Journal of Entrepreneurial Behavior & Research | C | 2 | 3.225 |
| 16 | Journal of Family Business Strategy | C | 2 | 2.605 |
| 17 | Creativity and Innovation Management | C | 2 | 2.015 |
| 18 | International Journal of Entrepreneurship and Innovation | C | 2 | – |
| 19 | International Journal of Entrepreneurship and Small Business | C | 2 | – |
| 20 | International Journal of Innovation and Technology Management | C | 2 | – |
| 21 | International Journal of Entrepreneurship and Innovation | C | 2 | – |
| 22 | Venture Capital: An International Journal of Entrepreneurial Finance | C | 2 | – |
| 23 | International Journal of Entrepreneurship and Innovation Management | C | 1 | – |
| 24 | Journal of Enterprising Culture | C | 1 | – |
| 25 | International Journal of Globalisation and Small Business | C | 1 | – |
| 26 | Journal of Entrepreneurship | C | 1 | – |
| 27 | Journal of International Entrepreneurship | C | 1 | – |
| 28 | Journal of Small Business and Entrepreneurship | C | 1 | – |
| 29 | Journal of Developmental Entrepreneurship | C | – | – |
| 30 | Journal of Entrepreneurial Finance and Business Ventures | C | – | – |
| 31 | Journal of Entrepreneurship Education | C | – | – |
| 32 | Journal of Family Business Management | C | – | – |
| 33 | Frontiers of Entrepreneurship Research | C | – | – |
| 35 | Journal of Research in Marketing and Entrepreneurship | C | – | – |
| 36 | Journal of Small Business Strategy | C | – | – |
| 37 | Zeitschrift für KMU und Entrepreneurship | C | – | – |
| 38 | Entrepreneurship Research Journal | – | 2 | 1.625 |
| 39 | Journal of Small Business and Enterprise Development | – | 2 | – |
| 40 | International Review of Entrepreneurship | – | 2 | – |
| 41 | Journal of Social Entrepreneurship | – | 2 | – |

**Table 4** Conversion table of leading academic journal rankings (based on Bouncken et al. 2015; amended/updated)

| VHB JQ3 | ABS | JCR IF |
|---|---|---|
| A+ | 4* | ≥5.0 |
| A | 4* | ≥3.5 |
| B | 3* | ≥2.5 |
| C | 2* | ≥1.5 |
| D | 1* | ≥0 |

## The process of writing a systematic literature review in entrepreneurship

Several authors presented a model for the different stages of an SLR (Briner and Denyer 2012; Frank and Hatak 2014; Jones and Gatrell 2014; Okoli 2015; Pittaway et al. 2014; Tranfield et al. 2004; Tranfield et al. 2003). While Tranfield et al. (2003) separate five stages in 10 steps, Frank and Hatak (2014) split them into six steps. Okoli (2015) explains four steps and Pittaway et al. (2014) uses three steps for an SLR. The main steps for all descriptions are always the same, namely: *Planning the review, Conducting the review* and *Reporting of the findings*. Some authors pay particular attention to the creation process, while others give more importance to reporting the results. We follow the three main steps and divide conducting the review into the identification and the synthesis of literature to underline the importance of these two steps in the process (Table 5).

### Planning the review

A good systematic review can be conducted to be published as an individual (standalone) paper and should be considered that way. However, it can further be part of a larger research project (e.g. a cumulative dissertation), where it is not only a published paper, but more importantly, a useful overview to show the importance of the main topic and to help identify further white spots the dissertation can search answers for. In both cases, it is important that the literature review is executed in a professional manner. Thus, the planning of the review article is of significant importance.

**Table 5** Process for SLR in Entrepreneurship (adapted from Tranfield et al. 2003, 2004)

| |
|---|
| Stage 1: Planning the review |
| - Identify the need |
| - Develop protocol |
| Stage 2: Identifying and evaluating studies |
| Stage 3: Extracting and synthesizing data |
| - Conducting data extraction |
| - Conducting data synthesis |
| Stage 4: Disseminate the review findings |

### Identify the need

In a first step, authors must identify if there is a need for an SLR. Therefore, they must become familiar with the literature. Check if there is already an SLR covering the topic being published or if this is the first of its kind. To do so, we suggest starting with literature searches on Google Scholar or potential databases that cover most of the Entrepreneurship literature. If there already is an SLR you must decide if different questions can be asked that can be answered with new research or if the existing research has been carried out badly. Both situations allow the author to think about writing new research. To answer the questions, the author has to be familiar with the methodology of SLR. This paper provides a first step to support you. The following list provides some topics and benefits SLRs deal with in general (Palmatier et al. 2018). The list is neither is extensive but not exhaustive. If the literature allows further contributions, the authors should take advantage of this.

- Create clear definitions by considering defining approaches of several authors from different perspectives,
- synthesize the existing literature and highlight important issues,
- point out irregular or special results,
- assess methodologies and results,
- take advantage of existing research to develop frameworks
- Highlight research gaps and potentially fruitful research directions

An SLR is basically created around a research question that it tries to answer. Independent of the purpose of the SLR, there is a general aim, hypothesis or question it addresses. This issue is built around the question of what we are looking for with the review article (Briner and Denyer 2012). There are different strategies, and in more mature fields, an SLR can handle very specific questions and be used for review articles that are strongly focused on depth than breadth to deal with specific types of research (qualitative, quantitative, case studies) only. The research question is a central part of the SLR and motivates the topic (Fisch and Block 2018). Most review articles in Entrepreneurship cover wide topics and provide a synthesis of general topics. They provide useful information for further contributions on the topic; however, we further want to encourage authors to conduct literature reviews that deal with the topic in an in-depth manner. They could help to cover and overcome the often-occurring definition problem in Entrepreneurship.

### Develop a review protocol

To ensure a transparent and high-quality process, authors have to create a review protocol at the beginning (Tranfield et al. 2003). The protocol is the basis for the ongoing research. It contains the whole process and therefore supports the methodology of the review article. The protocol outlines the parameters for the data search. This comprehensive protocol deals with the search strings, the databases, the criteria for including or excluding literature, quality criteria and so on (Pittaway et al. 2014). There are several worthy articles and books that deal with the topic of a review protocol. It is essential that you do not consider a review protocol to be set in stone, but rather that every change should be mentioned and protocoled.

Several authors encourage the writers of a literature review to search for all available literature, as books, conference papers and grey literature (Briner and Denyer 2012). However, for Entrepreneurship literature reviews, we would encourage authors to conduct their search mainly via online databases and for journal articles only (as the most "valuable" sources in research), as this search strategy helps to create a more transparent process that can be applied globally. There is an ongoing discussion today is about the use of *grey literature* (such as working papers, conference proceedings etc.) in systematic literature reviews, but we would advise against including those. Peer reviewed journal articles are checked through the academic process, while other literature is mostly unchecked (Podsakoff et al. 2005), which makes these supposedly "stronger" and thus more widely accepted as higher quality source (acknowledging the potential inherent publication bias, as without any question good research can also be published in other sources of publication as a journal). In research, we already see several systematic reviews that follow the process of including peer-reviewed articles only (e.g. Bouncken et al. 2015; Jones et al. 2011). For Entrepreneurship, following the vast majority of previous research in the field, we suggest concentrating on the main databases: ABI Inform/ProQuest, EBSCO/ Business Source Premier, JSTOR, MENDELEY, ScienceDirect, Scopus, SpringerLink, and Web of Science. Authors should use more than one database to cover most articles (Bramer et al. 2017). While Google Scholar can help to find full texts of papers and discover grey literature (Haddaway, 2015), and although it can also serve as an additional basis for cross-checks, it is not reproducible as the algorithm shows results for the author based on the prior searches and interactions (Gusenbauer and Haddaway 2019) and lists too many non-academic sources, and should therefore not be used for a systematic literature review.

Based on the keyword strings that are created for the review protocol, the author searches the databases. After a first search, a cross read through some articles helps to identify missing keywords for the search strings. Especially in Entrepreneurship, there is often more than one keyword for the same topic. For example, "Corporate Entrepreneurship" and "Intrapreneurship" deal with more or less the same topic, but they are completely different keywords. Thus, the author has to be careful to find all the necessary literature. To identify more keywords, consultations with experts in the topic are useful. Furthermore, educational literature helps to uncover synonyms. Another important question the author must answer is where to search. 'Only in Title' or 'in Title and abstract' or even 'in the text'. This decision can be a challenge, but is very important and should always be considered with the question of whether the scientists is writing a literature review in breadth or depth.

## Identifying and evaluating studies

The identification of studies for the review article is based on the topics handled previously. As already mentioned, we encourage Entrepreneurship authors to concentrate on electronic databases and peer reviewed journal articles only. This approach helps to ensure the highest standards of transparency. By contrast, there are authors that do not believe that a critical appraisal of a topic is possible with published journal articles only, and because of a publication bias, grey literature is needed (Briner and Denyer 2012). However, traditional reviews are criticized for subjective literature selection and quality appraisal (Denyer and Tranfield 2006). The use of grey literature

would open the SLR to this criticism. Petticrew and Roberts (2006) justify the search of grey literature, as in 2006 there existed cases where only a third of the literature finally used could be found in electronic databases. This should be reconsidered today as the availability of literature in electronic databases has increased dramatically. Furthermore, their process of searching for literature is highly complicated and difficult to be set into a transparent methodology. Through their encouragement of adding grey literature, they open the sciences to non-scientific publications. Even if the quality of these publications is evaluated by the authors themselves, it at least provides a more subjective filter than trusting into the academic processes from the beginning on where a double-blind review system is used to ensure high quality. So, we rather disagree with that view, and encourage authors to trust the journal reviewers and editors as this process is more objective than the exclusion on the literature reviewer's opinion.

Authors should not rate the publications by themselves as they can trust in the three main journal rankings available. For Entrepreneurship we suggest to only include literature that is published at least on level "C" of the VHB rating or equivalent (see Table 5). If there is some literature that is very suitable for the literature review but not published in a highly ranked journal, the author can give reason based on the quality of the article to add it into the bulk of reviewable literature. However, these additional publications should be added carefully and evaluated deeply as other journals are likely to have already rejected the article. The combination of the three major journal rankings are a transparent and reliable way of evaluating studies.

After excluding studies that do not meet the quality criteria, the authors must start by reading through the titles of the remaining studies. Often the title alone can reveal whether a study fits the review criteria mentioned in the review protocol or not. After a first read through the titles and this exclusion round, another one can be conducted based on the abstract and the research question the author wants to answer. This is especially relevant for research questions that want to dig deep in a narrow niche the abstract can help to further exclude articles. If a title only search already creates enough literature to allow for a synthesis this will be enough. However, if there needs to be more literature, the search should be expanded to the abstract too. A search in the whole text may be useful to uncover the keyword in papers that are only loosely related to the main topic. Therefore, we suggest not to search whole texts. After excluding all articles that does not provide any useful information to answer the question or do not fit the quality criteria, the data extraction can start.

## Extracting and synthesizing data

### Conducting data extraction

For an SLR, the data extraction has to be systematic and transparent. Therefore, the author has to describe the data for the extraction in the review protocol and create an extraction sheet at the beginning. As mentioned before, the protocol and the sheet can be adapted if important issue arises during the extraction. However, the change of both tools must be noted. During this process, the author could create a "data extraction bias" by different judgement of the studies (Petticrew and Roberts 2006). To overcome the bias, prevent missing important data and create a higher level of objectivity, more than one author should conduct the data extraction (Rousseau et al. 2008).

The actual data extraction depends on the studies investigated and the research question. Tables are considered a useful support to create an overview and a transparent matrix for the ongoing synthesis (Petticrew and Roberts 2006). The table should contain all the necessary information for the synthesis and every paper reviewed needs to be outlined there. A first step necessary to allow for a concept centric synthesis is the organization of papers by the author. There is no all-embracing rule for this organization as it is dependent on the research question and the nature of the papers being analyzed. The categories that the author wants to sort papers into should be considered in the table. For this section, we would like to mention Newbert (2007) as an example of how to use tables for the data extraction and further on for the synthesis.

### Conducting data synthesis

The data synthesis is one of the most important steps in writing an SLR. While an increasing number of SLRs replace traditional reviews in journals, the main reason for rejection is a lack in the quality of the review article. SLRs have to analyze and compare existing literature instead of just summarizing it (Jones and Gatrell 2014). Increasing numbers of papers, constructs and methodologies create a more difficult environment to conduct a professional synthesis of the topic which often leads to more specific and deeper review articles on a limited topic (Palmatier et al. 2018).

To synthesize the data, authors can follow different strategies. For an SLR, it is important to concentrate on concepts and not on authors and their studies. This concept centric writing style should be constituted in the microstructure of the paper (Fisch and Block 2018). An author- centric approach is not suitable for a successful literature review, as this strategy is more likely to end in a summary than a synthesis (Webster and Watson 2002).

For the synthesis of the literature, an objective view of the author is necessary. It is of major importance that the author not only analyses the results of the study but also the methodology to identify problems and make comparisons to other studies results. If not, there is a high probability that the synthesis is biased as results from two studies can be contradicting, based on the quality of the methodology used (Light and Smith 1971; Sutton et al. 2000). The synthesis of the literature review basically depends on the question and the goals of the review article. Practical tips for the implementation of a particularly successful synthesis therefore largely depend on the circumstances. So, we exclusively refer to successful and good SLRs here (e.g. Liñán and Fayolle 2015; Newbert 2007; Pittaway and Cope 2007; Stephan 2018).

### Disseminating the review findings

A systematic review like the SLR is located high on the level of evidence hierarchy. While a lot of empirical papers deal with specific situations, the review article combines and synthesizes a lot of these studies and helps to create evidence-informed information. So, we can say individual empirical studies are more likely to be considered in research, but they still need to address the problems the practitioners face. An SLR, on the other hand, is strongly tied to practitioners too as the findings are more general. So, in Entrepreneurship, the target groups of an SLR are the entrepreneurs and or managers and fellow researchers. Bem (1995) further supports this statement by dealing with the

review topic in *Psychological Bulletin*. He states that publications reviewing the literature should be understandable by more than pure experts in the topic, therefore he provides a range of writing techniques that address a broad audience.

## General settings

### Reviewing team

In research, few articles are single authored today. A strong team increases the value of a paper, as the team can be seen as pre-reviewers. For an SLR, this issue also applies and we encourage authors to team up. Jones and Gatrell (2014) provide a section on how a team could be put together. An opening pieces of advice suggests heterogeneity by age, constituted through a balance between more experienced scholars and younger researchers – a combination of Ph.D. graduates and their supervisors for example (Akinci and Sadler-Smith 2012). SLRs are often described to be interdisciplinary as they search for and synthesize literature that is founded in different disciplines. Therefore, the suggestion is also to create interdisciplinary teams to overcome potential bias when not understanding the other disciplines that are included in the SLR (Jones and Gatrell 2014).

### Structure of a review

The general structure of a review article is very similar to an empirical article. It starts with an introduction to motivate the topic and deals with its contributions to research and practice (Webster and Watson 2002). The article then deals with the methodology (Denyer and Neely 2004) and the synthesis of the reviewed literature. This is followed by a discussion section and a conclusion (Fisch and Block 2018). However, this structure is not compulsory and can be adapted. In Entrepreneurship SLRs, authors often include an extra chapter after the introduction to give some background information about the theoretical underpinnings of the topic before dealing with the methodology (e.g., Dorn et al. 2016; Hakala 2011; Stephan 2018). Thus, they try to organize the paper very similarly to an empirical paper. This first theoretical background does not need to exist independently to the literature that is used for the systematic review later on. Rather, it can be seen as a short traditional review to motivate the research question of the SLR.

From a microstructure perspective, the organization of the paper within the chapters is more difficult than in empirical papers (Bem 1995). The microstructure can be very diverse among different papers and is always driven by the main research question and the synthesis of the authors.

### Number of articles

An SLR is dependent on existing data, i.e. previous research on the topic. We have already dealt with the issue of "when is an SLR suitable". However, another important question is how many articles a literature review should contain. Authors answered that question quite differently ranging from several hundred articles to a few dozen (Frank and Hatak 2014). At this point, we differentiate between SLRs in immature or mature

research fields. Depending on the maturity, the SLR can be used to follow different aims.

In a less mature field, the number of available articles is limited and more scattered as a lot of research questions remain unanswered. At this point, an SLR can help to establish a new theory on the basis of existing articles (Frank and Hatak 2014). Another reason to conduct an SLR in an immature field is to point out missing data and call for empirical research at the right point of time (Petticrew and Roberts 2006). An early SLR can further provide a better-established definition and understanding of the research field, so researchers can start to work on the field with a general understanding. In immature fields, review articles are less hypothesis or research question driven, and more strongly focused on synthesizing the basic foundations of the field and provide valuable insights. The evidence-based approach is not the focus there.

In a more mature field, the pattern of a literature review can change and is more evidence driven. It follows research questions and tries to answer hypotheses. At this stage, the number of published articles is higher and therefore more topics are investigated. A major question arising at this point is how deep and broad the SLR should be (Fisch and Block 2018).

For both scenarios, immature and mature, it is always important to tie the number of articles to the aim of the SLR. However, in immature fields and for theory creation, a lower number of articles is plausible (Frank and Hatak 2014). A low number of articles can also be justified for an SLR that covers a very specific topic in a mature field.

Furthermore, we want to draw attention to the opposite case of a very high number of articles being available. Although modern technology enables researchers to choose, collect and analyze publications quicker for the synthesis of the articles, a high level of resources is needed, as the quality of an SLR is highly dependent on this step. Claiming an SLR on hundreds of articles is not enough if the important connections among these articles are not identified. The author should limit the topic or be more selective and maintain the quality. We finish our article with a list of concrete tips which might be helpful for conducting your own SLR (see Table 6).

## Conclusion

In recent years, SLRs have become more popular in Entrepreneurship literature, and an increasing number of journals started to publish literature reviews as standalone articles. However, one of the main reasons review articles are regularly rejected is the lack of quality of their synthesis. This issue can only be overcome by better guidance of the authors. Our paper shows the main differences of a traditional literature review and an SLR. When reading about systematic reviews, we can see that this methodology has changed a lot in recent years because of technical innovations and software support. Dated publications about the topic are clearly written with a limitation in mind due to software logic. The authors evidently did not trust these databases and their search functions. They also did not have the possibility to download literature to the same extent as can be done today. Therefore, we suggested some steps to be conducted differently today.

For the literature search, we strongly believe that authors can trust the major online databases and take the literature from there. We do not foresee a lack of literature by

**Table 6** Concrete tips for your SLR

1. Find an area of research of scholarly (and potentially also practical) interest.
2. Undertake an initial short literature search to gain an overview on the topic. Google Scholar can be used for this. Find keywords and synonyms. Actively search for already existing literature reviews.
3. Find questions in the area of research that can be answered with an SLR and define the purpose of your review article: answer questions, map the literature (in case of dissertation project).
4. The review of literature (i.e., of the state-of-the-art of research in your field) leads (as a result) to the research question/empirical model.
5. Create a review and data extraction protocol based on the purpose of your SLR. Excel is a useful tool to do so.
6. Define keywords and search strings. Get feedback on the search strings so you do not miss any synonym or relevant literature.
7. For the SLR, you need to at least try to find EVERY academic source on your topic by using the academic literature databases (such as EBSCO/Business Source Complete, Web of Science, ScienceDirect, ABI--Inform/ProQuest, JSTOR etc.). Always use more than one, if not all to which you have access! Do not use Google Scholar for this step of the process; concentrate on academic journal articles ONLY, since those are the "best" sources because they are *peer reviewed* (by other scholars).
8. Be transparent: Note the date you searched for the literature and record the number of articles you found in each database.
9. If you find a potentially relevant article in your literature search, but to which you do not have access over all used databases in full text, you need to utilize the following possibilities to get this text: ResearchGate.net, contact the corresponding author, check if friends at other institutions might have increased access to databases.
10. Check the quality of the articles found through the journal rankings. As a rule of thumb, everything from "A" to "C" is more than acceptable. Use the journal ranking transformation table in this paper to decide which article should be used or not for the SLR.
11. Read through abstracts to decide if the article can really help to answer the research question.
12. Extract data into your data extraction form. Record changes on the data extraction form. Provide a transparent and objective process.
13. Synthesize the data you extracted in the best way. Concentrate on concepts and not authors. You should not do a pure summary of the content. Keep in mind that this section is the reason for most desk rejections. Use tables and figures to synthesize the data.
14. Depending on the time since the literature search in the databases, the databases should be searched again for new literature before submission.
15. Write your article and submit it to your target journal.
16. Hope for reviewers who know the field and/or method and provide valuable feedback which helps to increase the quality of your article in the revision rounds, so that it can finally get accepted for publication!

only searching online databases in the Entrepreneurship domain. However, it should be clear that a database search cannot be limited to a single database. While in former times, the hard part was to detect and gain access to literature, today it is to define the perfect keywords and strings. Pittaway and Cope (2007) defined 27 keywords and 10 final search strings to discover the literature needed. Furthermore, the ongoing trend of "open access" and the reduction of pay-walls provides easier access to the literature (Jones and Gatrell 2014). In contrast, this overflow of literature increases the importance of perfectly defined database search strings, as otherwise the author is faced with hundreds of potentially irrelevant articles.

One of the major differences among traditional reviews and the SLR is the higher level of objectivity and transparency that the SLR provides. While some authors argue

that the authors should set the quality criteria, we argued to trust into academic processes and the results of the three main journal rankings. The use of the journal rankings to set quality thresholds creates a higher level of objectivity as the deletion of articles is always transparent and not driven by the opinion of the author.

Literature reviews are also a foundation for dissertations and other research proposals (Heath and Tynan 2010). An SLR opens the topic and maps the literature. While an SLR as a standalone paper allows to answer a research question, the main purpose of a SLR for a dissertation is to uncover potential research questions and hypotheses for the further research on this topic. Therefore, they are following two different aims that result in slight differences. However, the basic foundations for both purposes are the same: being systematic and transparent.

A good literature review is generally based on the abilities of the author or the team that conducts the review article. However, technology plays a role of increasing importance today. The Internet and online databases have already helped to create a faster process for SLRs in recent years and allow paper collection to be more transparent. Open access trends further contribute to the possibility of better literature reviews through easier availability. So, compared with early processes of literature reviews, current processes are less time consuming and allow for a higher level of transparency and reproducibility. Furthermore, journal rankings allow authors to set a first quality criterion for the articles gathered and provide the possibility of a transparent article selection and exclusion. However, for the step of data extraction and synthesis, the author is a potential factor for confirmation bias. Furthermore, these are areas where most SLRs suffer from a lack of quality (Jones and Gatrell 2014). Authors do read a lot of papers on a specific topic and can never be sure that they have not overlooked important issues in the papers. Furthermore, the maximum number of papers they can look through is limited by time and financial resources. Today, software can assist to overcome some of the limitations of an SLR. Rauch (2019) showed some examples where software was used to analyze qualitative data (e.g. Kaminski and Hopp 2019; Short et al. 2010; von Bloh et al. 2019). Automated data analysis would lead to several advantages for literature reviews. In particular, this analysis can help to analyze a greater volume of literature in a short period of time. So, the maximum number of articles for an SLR is not capped by the time resources of the research team. Furthermore, computer-based analysis can be more objective than comparable human analysis. To ensure this advantage it is important that transparency is provided. So, authors should use the same algorithms over time, or contribute their source code publicly on GitHub.

The use of algorithms is so far mainly used only to analyze non-academic texts. To our knowledge, systematic reviews in Entrepreneurship that are created with the support of algorithms do not exist yet. Natural Language Processing is needed in order to analyze texts. This analytical technique is used to make natural language processable for computers (Yim et al. 2016). Nevertheless, analytic software-supported SLRs can also result in review articles with a lack of transparency and non-reproducibility. Accumulated research can influence the algorithm, especially if it is interested in falsifying the results. For transparency issues it is important that algorithms that analyzed the literature are available to other researchers too.

**Acknowledgments** We would like to thank Rodrigo Isidor, Paul Jones, Patrycja Klimas, Thomas Niemand and the two anonymous reviewers for their valuable comments on the paper and its revision.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Akinci, C., & Sadler-Smith, E. (2012). Intuition in management research: A historical review. *International Journal of Management Reviews, 14*(1), 104–122.

Armitage, A., & Keeble-Allen, D. Undertaking a structured literature review or structuring a literature review: tales from the field. In *Proceedings of the 7th European Conference on Research Methodology for Business and Management Studies: ECRM2008, Regent's College, London, 2008* (pp. 35).

Bem, D. J. (1995). Writing a review article for psychological bulletin. *Psychological Bulletin,* 118(2), 172.

von Bloh, J., Broekel, T., Özgun, B., & Sternberg, R. (2019). New (s) data for entrepreneurship research? *An innovative approach to use big data on media coverage. Small Business Economics*, 1–22.

Bouncken, R. B., Gast, J., Kraus, S., & Bogers, M. (2015). Coopetition: A systematic review, synthesis, and future research directions *Review of Managerial Science, 9*(3), 24.

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews, 6*(1), 245.

Briner, R. B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In D. M. Rousseau (Ed.), *Handbook of evidence-based management: Companies, classrooms and research* (p. 17). New York: Oxford University Press.

Calabrò, A., Vecchiarini, M., Gast, J., Campopiano, G., De Massis, A., & Kraus, S. (2019). Innovation in family firms: A systematic literature review and guidance for future research. *International Journal of Management Reviews, 21*(3), 317–355.

Davies, H. T., & Nutley, S. M. (1999). The rise and rise of evidence in health care. *Public money and management, 19*(1), 9–16.

De Bakker, F. G., Groenewegen, P., & Den Hond, F. (2005). A bibliometric analysis of 30 years of research and theory on corporate social responsibility and corporate social performance. *Business & Society, 44*(3), 283–317.

Denyer, D., & Neely, A. (2004). Introduction to special issue: Innovation and productivity performance in the UK. *International Journal of Management Reviews, 5*(3–4), 131–135.

Denyer, D., & Tranfield, D. (2006). Using qualitative research synthesis to build an actionable knowledge base. *Management Decision, 44*(2), 213–227.

Dorn, S., Schweiger, B., & Albers, S. (2016). Levels, phases and themes of coopetition: A systematic literature review and research agenda. *European Management Journal, 34*(5), 484–500.

Ferreira, J. J. M., Fernandes, C. I., & Kraus, S. (2019). Entrepreneurship research: Mapping intellectual structures and research trends. [journal article]. *Review of Managerial Science, 13*(1), 181–205.

Fisch, C., & Block, J. (2018). Six tips for your (systematic) literature review in business and management research. *Management Review Quarterly, 68*(3).

Frank, H., & Hatak, I. (2014). Doing a research literature review. In A. Fayolle & M. Wright (Eds.), *How to get published in the best entrepreneurship journals: A guide to steer your academic career* (p. 23). Cheltenham: Edward Elgar Publishing.

Gusenbauer, M., & Haddaway, N. R. (2019). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google scholar, PubMed and 26 other resources. *Research Synthesis Methods*.

Hakala, H. (2011). Strategic orientations in management literature: Three approaches to understanding the interaction between market, technology, entrepreneurial and learning orientations. *International Journal of Management Reviews, 13*(2), 199–217.

Hart, C. (1998). *Doing a literature review*. London: Sage Publications.

Heath, M., & Tynan, C. (2010). Crafting a research proposal. *The Marketing Review, 10*(2), 147–168.

Hodgkinson, G. P., & Ford, J. K. (2014). Narrative, meta-analytic, and systematic reviews: What are the differences and why do they matter? *Journal of Organizational Behavior, 35*(S1), S1–S5.
Hodgkinson, G. P., & Ford, J. K. (2015). What makes excellent literature reviews excellent? A clarification of some common mistakes and misconceptions. *Journal of Organizational Behavior, 36*(S1), S1–S5.
Jones, O., & Gatrell, C. (2014). Editorial: The future of writing and reviewing for IJMR. *International Journal of Management Reviews, 16*, 249–264.
Jones, M. V., Coviello, N., & Tang, Y. K. (2011). International entrepreneurship research (1989–2009): A domain ontology and thematic analysis. *Journal of Business Venturing, 26*(6), 632–659.
Kaminski, J. C., & Hopp, C. (2019). Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. *Small Business Economics*, 1–23.
Knopf, J. W. (2006). Doing a literature review. *PS: Political Science and Politcs, 39*(1), 5.
Light, R., & Smith, P. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review, 41*(4), 429–471.
Liñán, F., & Fayolle, A. (2015). A systematic literature review on entrepreneurial intentions: Citation, thematic analyses, and research agenda. *International Entrepreneurship and Management Journal, 11*(4), 907–933.
Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737–738.
Mulrow, C. D. (1994). Systematic reviews: Rationale for systematic reviews. *British Medical Journal, 309*, 597–599.
Newbert, S. L. (2007). Empirical research on the resource-based view of the firm: An assessment and suggestions for future research. *Strategic Management Journal, 28*(2), 121–146.
Oakley, A. (2002). Social science and evidence-based everything: The case of education. *Education Review, 54*, 277–286.
Ohlsson, A. (1994). Systematic reviews-theory and practice. *Scandinavian Journal of Clinical and Laboratory Investigation, 54*(sup219), 25–32.
Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems, 31*(37).
Palmatier, R. W., Houston, M. B., & Hulland, J. (2018). Review articles: Purpose, process, and structure. Springer.
Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell Publishing.
Pittaway, L., & Cope, J. (2007). Entrepreneurship education: A systematic review of the evidence. *International Small Business Journal, 25*(5), 479–510.
Pittaway, L., Holt, R., & Broad, J. (Eds.). (2014). *Synthesising knowledge in entrepreneurship research: The role of systematic literature reviews* (*Handbook of research on small business and entrepreneurship*). London: Edward Elgar.
Podsakoff, P. M., MacKenzie, S. B., Bachrach, D. G., & Podsakoff, N. P. (2005). The influence of management journals in the 1980s and 1990s. *Strategic Management Journal, 26*(5), 473–488.
Rauch, A. (2019). Opportunities and threats on reviewing entrepreneurship theory and practice. *Entrepreneurship Theory and Practice*, 1–14.
Rousseau, D. M., Manning, J., & Denyer, D. (2008). Chapter 11: Evidence in management and organizational science: Assembling the Field's full weight of.
Rowley, J., & Slack, F. (2004). Conducting a literature review. *Management Research News, 27*(6), 31–39.
Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods, 13*(2), 320–347.
Stephan, U. (2018). Entrepreneurs' mental health and well-being: A review and research agenda. *Academy of Management Perspectives, 32*(3), 290–322.
Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research* (Vol. 348): Wiley Chichester.
Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*, 207–222.
Tranfield, D., Denyer, D., Marcos, J., & Burr, M. (2004). Co-producing management knowledge. *Management Decision, 42*(3/4), 375–386.
Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly, xiii-xxiii.*
Yim, W. W., Yetisgen, M., Harris, W. P., & Kwan, S. W. (2016). Natural language processing in oncology: A review. *JAMA Oncology, 2*(6), 797–804.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Scientometrics (2021) 126:871–906
https://doi.org/10.1007/s11192-020-03690-4

# Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations

**Alberto Martín-Martín**[1] · **Mike Thelwall**[2] · **Enrique Orduna-Malea**[3] · **Emilio Delgado López-Cózar**[1]

Received: 30 April 2020 / Published online: 21 September 2020
© Akadémiai Kiadó, Budapest, Hungary 2020, corrected publication 2020

**Abstract**
New sources of citation data have recently become available, such as Microsoft Academic, Dimensions, and the OpenCitations Index of CrossRef open DOI-to-DOI citations (COCI). Although these have been compared to the Web of Science Core Collection (WoS), Scopus, or Google Scholar, there is no systematic evidence of their differences across subject categories. In response, this paper investigates 3,073,351 citations found by these six data sources to 2,515 English-language highly-cited documents published in 2006 from 252 subject categories, expanding and updating the largest previous study. Google Scholar found 88% of all citations, many of which were not found by the other sources, and nearly all citations found by the remaining sources (89–94%). A similar pattern held within most subject categories. Microsoft Academic is the second largest overall (60% of all citations), including 82% of Scopus citations and 86% of WoS citations. In most categories, Microsoft Academic found more citations than Scopus and WoS (182 and 223 subject categories, respectively), but had coverage gaps in some areas, such as Physics and some Humanities categories. After Scopus, Dimensions is fourth largest (54% of all citations), including 84% of Scopus citations and 88% of WoS citations. It found more citations than Scopus in 36 categories, more than WoS in 185, and displays some coverage gaps, especially in the Humanities. Following WoS, COCI is the smallest, with 28% of all citations. Google Scholar is still the most comprehensive source. In many subject categories Microsoft Academic and Dimensions are good alternatives to Scopus and WoS in terms of coverage.

**Keywords** Google Scholar · Microsoft Academic · Scopus · Dimensions · Web of Science · OpenCitations · COCI · CrossRef · Coverage · Citations · Bibliometrics · Citation analysis · Bibliographic databases · Literature reviews

✉ Alberto Martín-Martín
albertomartin@ugr.es

[1] Facultad de Comunicación y Documentación, Universidad de Granada, Granada, Spain

[2] Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton, UK

[3] Universitat Politècnica de València, Valencia, Spain

## Introduction

### Timeline

The first scientific citation indexes were developed by the Institute for Scientific Information (ISI). The Science Citation Index (SCI) was introduced in 1964, and was later joined by the Social Sciences Citation Index (1973) and the Arts & Humanities Citation Index (1978). In 1997, these citation indexes were moved online under the name "Web of Science". Recently, these citation indexes, along with some new ones such as the Conference Proceedings Citation Index, the Book Citation Index, and the Emerging Sources Citation Index, were rebranded as the "Web of Science Core Collection" (from now on, WoS). The availability of this data was essential to the development of quantitative studies of science as a field of study (Birkle et al. 2020).

In November 2004, two new academic bibliographic data sources that contained citation data were launched. Like WoS, Elsevier's Scopus is a subscription-based database with a selective approach to document indexing (documents from a pre-selected list of publications). A few weeks after Scopus, the search engine Google Scholar was launched. Unlike WoS and Scopus, Google Scholar follows an inclusive and automated approach, indexing any seemingly academic document that its crawlers can find and access on the web, including those behind paywalls through agreements with their publishers (Van Noorden 2014). Additionally, Google Scholar is free to access, allowing users to access a comprehensive and multidisciplinary citation index without charge.

In 2006, Microsoft launched Microsoft Academic Search but retired it in 2012[1] (Orduña-Malea et al. 2014). In 2016, Microsoft launched a new platform called Microsoft Academic, based on Bing's web crawling infrastructure. Like Google Scholar, Microsoft Academic is a free academic search engine, but unlike Google Scholar, Microsoft Academic facilitates bulk access to its data via an Applications Programming Interface (API) (Wang et al. 2020).

In 2018, Digital Science launched the Dimensions database (Hook et al. 2018). Dimensions uses a freemium model in which the basic search and browsing functionalities are free, but advanced functionalities, such as API access, require payment. This fee can be waived for non-commercial research projects.

Also in 2018, the organization OpenCitations, dedicated to developing an open research infrastructure, released the first version of its COCI dataset (OpenCitations Index of CrossRef open DOI-to-DOI citations). The citation data in COCI comes from the lists of references openly available in CrossRef (Heibi et al. 2019). Until 2017, most publishers did not make these references public, but the Initiative for Open Citations (I4OC), launched in April 2017, has since convinced many publishers to do so. The rationale is that citation data should be considered a part of the commons and should not be only on the hands of commercial actors (Shotton 2013, 2018). At the time of writing, 59% of the 47.6 million articles with references deposited with CrossRef have their references open.[2] However, some large publishers, such as Elsevier, the American Chemical Society, and IEEE have not yet agreed to opening their lists of references. Thus, COCI's only partially reflects the

[1] https://web.archive.org/web/20170105184616/https:/academic.microsoft.com/FAQ

[2] https://i4oc.org/.

citation relationships of documents recorded in CrossRef, which now covers over 106 million records (Hendricks et al. 2020).

The new bibliographic data sources are changing the landscape of literature search and bibliometric analyses. The openly available data in Microsoft Academic Graph (MAG) has been integrated into other platforms, significantly increasing their coverage (Semantic Scholar, Lens.org). There are still some reuse limitations, such as that the current license of MAG (ODC-BY) requires attribution, which apparently precludes it from being able to be integrated into COCI (which uses a CC0 public domain license). This openness is nevertheless an advance on the previous situation, in which most citation data was either not freely accessible (WoS, Scopus), or free but with significant access restrictions (Google Scholar). At this point, citation data is starting to become ubiquitous, and even owners of closed bibliographic sources, such as Scopus, are beginning to offer researchers options to access their data for free.[3]

Other citation indexes have been developed within various academic platforms, but these are not analysed in this study, for various reasons:

- CiteSeerX,[4] from Penn State University, indexes documents in the public web and not those that are only found behind paywalls (Wu et al. 2019).
- ResearchGate[5] generates its own citation index based on the full text documents that its crawler finds on the Web and those that its users upload to the platform. However, the platform does not offer a way to extract data in bulk, and it is difficult to use web scraping to obtain data because the complete list of citations to an article cannot be easily displayed.
- Lens.org[6] integrates coverage from Microsoft Academic, CrossRef, PubMed, and a number of Patent datasets. It is not included in this analysis because two of its main sources (Microsoft Academic and CrossRef) are already included.
- Semantic Scholar[7] originally focused on Computer Science and Engineering. Later it expanded to include Biomedicine, and recently it has integrated multidisciplinary coverage from Microsoft Academic (which is also the reason why we decided not to analyse it).
- There are also several regional or subject-specific citation indexes, which only index documents published by journals and/or researchers who work in a specific country or region, or in specific topics. Given their specific scope these are not easily comparable to sources with a worldwide and/or multidisciplinary coverage.

## Previous analyses of coverage

Document coverage varies across data sources (Ortega 2014), and studies that analyse differences in coverage can inform prospective users about the comprehensiveness of each database in different subject areas. For citation indexes, greater coverage should equate to higher citation counts for documents, if citations can be extracted from all documents.

[3] https://www.elsevier.com/icsr/icsrlab.
[4] https://citeseerx.ist.psu.edu/index.
[5] https://www.researchgate.net/.
[6] https://www.lens.org/.
[7] https://www.semanticscholar.org/.

Coverage is not the only relevant aspect that should be considered when deciding which data source should be used for a specific information need (e.g., literature search, data for bibliometric analyses). Other aspects such as functionalities to search, analyse, and export data, as well as transparency and cost, are also relevant, but not analysed here. Some of these aspects are analysed by Gusenbauer and Haddaway (2020).

### The veterans: WoS, Scopus, and Google Scholar

As the longest-running platforms, many studies have analysed the differences in coverage and citation data between WoS, Scopus, and Google Scholar. WoS covers over 75 million records in its Core Collection (which includes its main citation indexes), and up to 155 million records when other regional and subject-specific citation indexes are included (Birkle et al. 2020). Scopus claims to cover over 76 million records (Baas et al. 2020). Google Scholar does not disclose official figures about its coverage (Van Noorden 2014), but the most recent independent studies have estimated that it covers well over 300 million records (Delgado López-Cózar et al. 2019; Gusenbauer 2018). At this point most studies agree that Google Scholar has a more comprehensive coverage than Scopus and WoS, and includes the great majority of the documents that they cover. However, the relatively low quality of the metadata available in Google Scholar and the difficulty to extract it make it challenging to use Google Scholar data in bibliometric analyses (Delgado López-Cózar et al. 2019; Halevi et al. 2017; Harzing 2016a, b; Harzing and Alakangas 2016; Martín-Martín et al. 2018; Moed et al. 2016).

### Microsoft academic

Microsoft Academic has been recently reported to cover over 225 million publications (Wang et al. 2020). Harzing carried out an analysis of her own publication record and the publication records of 145 academics in five broad disciplinary areas (Harzing 2016a, b; Harzing and Alakangas 2017a, b). Microsoft Academic found more of her own publications than Scopus or WoS. For the sample of publications by 145 academics, Microsoft Academic provided higher citation counts than both Scopus or WoS in Engineering, Social Sciences, and the Humanities, and similar figures in Life Sciences and Sciences. Google Scholar reported the highest citation counts in all disciplines.

Hug and Brändle (2017) also analysed the coverage of Microsoft Academic and compared it to Scopus and WoS. Based on publications included in the repository of the University of Zurich as a case study, Microsoft Academic had wider coverage of non-article documents than Scopus and WoS, while Scopus had a slightly lower coverage of journal articles than Microsoft Academic. Microsoft Academic showed similar biases to Scopus and WoS against non-English publications and publications in the Humanities. Haunschild et al. (2018) analysed a subset of the same sample used in the previous study (25,539 papers also covered by WoS) and found that 11% had no associated cited references in Microsoft Academic, while in WoS the same papers had associated cited references. However, for publications with less than 50 associated references in WoS (24,788) the concordance correlation coefficient applied to the number of references found by each source was 0.68, indicating a strong tendency for them both to report the same number.

Thelwall (2017) analysed the citation counts of 172,752 articles in 29 large journals from various disciplines, and compared them to Scopus citation counts and Mendeley reader counts. For articles published between 2007 and 2017, Microsoft Academic found

slightly more citations than Scopus overall, and significantly more than Scopus for documents published in 2017. In subsequent studies, Thelwall (2018a) found that Microsoft Academic did find earlier citations to recently published articles when compared to Scopus. Kousha and Thelwall (2018) studied the coverage and citation counts of books in Microsoft Academic and Google Books by analysing a sample of book records extracted from the Book Citation Index (BKCI) in WoS. They found 60% of the books in their sample overall, but this percentage was lower in some categories of the Humanities and Social Sciences. Citation counts in Microsoft Academic were higher than in BKCI in 9 out of 17 fields during 2013–2016. Kousha et al. (2018) analysed whether Microsoft Academic was able to find early citations of in-press articles using a sample of 65,000 in-press articles from 2016 to 2017, and found that Microsoft Academic was able to find 2–5 times as many citations as Scopus. This was mostly because Microsoft Academic (like Google Scholar) merges preprints (and the citations these receive) with their subsequent in-press versions, and because Microsoft Academic covers repositories such as arXiv.

Visser et al. (2020) carried out a large-scale comparison of WoS, Scopus, Dimensions, Microsoft Academic, and CrossRef by matching the entire collection of documents in each source. They found that Microsoft Academic was the source with the largest coverage overall, and the one with the higher overlap with Scopus documents (81% of Scopus documents were found in Microsoft Academic). Some of the documents in Microsoft Academic were not of a scientific nature. Microsoft Academic was not able to detect 12.7% of the citations found by Scopus after adjusting for coverage differences.

### Dimensions

Dimensions covers over 105 million publications, as well as other kinds of records such as grants data, clinical trials, patents, and policy documents (Herzog et al. 2020).

Orduña-Malea & Delgado-López-Cózar (2018) analysed several small samples of journals, documents and authors in the field of Library & Information Science using Dimensions, and compared the data to Scopus and Google Scholar. Dimensions provided slightly lower citation counts than Scopus. Thelwall (2018c) analysed a random sample of 10,000 Scopus articles from 2012, finding that Dimensions covered the great majority of articles with a DOI (97%) and high correlations between citation counts in the two sources (median of 0.96 across narrow subject categories).

Harzing (2019) analysed coverage of Dimensions and CrossRef, and compared it to the coverage in WoS, Scopus, Google Scholar, and Microsoft Academic using her own publication and citation record, as well as that of six top journals in Business & Economics. CrossRef and Dimensions had similar or better coverage of publications, and similar citation counts to those in WoS and Scopus, but still substantively lower than Google Scholar and Microsoft Academic.

Visser et al. (2020) found that Dimensions had a substantially higher coverage than Scopus and WoS, which heavily relied on data from CrossRef. After computing the overlap in coverage between Dimensions and Scopus, they found that overall, Dimensions covered 78% of the documents available in Scopus (35.1 million out of 44.9 million documents in Scopus). They also analysed the accuracy and completeness of citation links, finding that, after adjusting for coverage differences, there were 489.7 million citations found by both sources (percentage of full overlap: 83%), 73.2 million only found by Scopus, and 25.8 million only found by Dimensions.

### COCI

COCI has detected over 624 million citation relationships involving over 53 million documents (Peroni and Shotton 2020). The citations recorded in this source are only a fraction of the citations that have actually occurred among the documents covered by CrossRef, because some publishers that deposit lists of references or CrossRef have not agreed to make them available, and other publishers and preprint servers do not deposit any references in CrossRef or do it only for some document types (Shotton 2018; van Eck et al. 2018). Huang et al. (2020) used citation data from COCI and bibliographic data from WoS, Scopus and Microsoft Academic to test the robustness of university rankings created with these different sources, and concluded that despite its lack of comprehensiveness COCI is already a viable data source for cross comparisons at the system level.

### Objective

The citation index coverage studies published so far have analysed a heterogeneous variety of samples of documents, disciplines, and data sources. In response, this paper reports a systematic comparison of coverage of six data sources (Google Scholar, Microsoft Academic, Scopus, Dimensions, WoS, and COCI[8]) across 252 subject categories using a relatively large sample of citations. This allows comparisons across a large number of disciplines for the most widely used bibliographic data sources. This study expands and updates a previous analysis of Google Scholar, Scopus and WoS (Martín-Martín et al. 2018). The main research question that drives this is investigation is:

*RQ* How much overlap is there between Google Scholar, Microsoft Academic, Scopus, Dimensions, WoS, and COCI in the citations that they find to academic documents and does this vary by subject?

## Methods

### Direct coverage comparison versus comparison of citations

The most direct method to compare document coverage across different data sources would be to obtain a complete list of all documents covered by each source, match the documents across databases, and report the size of the overlaps (Visser et al. 2020). This is not possible here because of access restrictions. For example, Scopus and WoS charge for this kind of access and Google Scholar does not share its database.

Because of these restrictions, studies analysing coverage differences across bibliographic data sources often use an alternative method: they select a seed sample of documents that are known to be covered by all the data sources under analysis, and then they compare the list of citing documents that each data source is able to find for each of the seed documents (Martín-Martín et al. 2018). The rationale of this method is that if data source *A* is not able to find a citation that data source *B* has found, the reason must be that

[8] In the case of COCI, the results cannot reflect the full coverage of CrossRef given the incomplete availability of reference lists in this source. Nevertheless, including it in the analysis will inform us of what proportion of citations are currently available in the public domain.

the citing document is not covered by data source *A*. This assumes that all data sources are equally effective in detecting citation relationships. In fact, each data source has its own (usually secret) citation detection algorithms, and small discrepancies in citation data across databases exist even when removing the factor of differences in coverage (van Eck and Waltman 2019; Visser et al. 2020). Furthermore, it is known that bibliographic databases do not always have access to cited reference lists for all the documents they cover, which also affects the citations they can detect. For example, reference lists are only available in a fraction of the documents indexed in CrossRef, so an analysis of the citations detected in this source does not accurately reflect the true size of the bibliographic database. Other sources, especially academic search engines, are also affected by this issue to some degree.[9] Lastly, academic documents that do not cite and are not cited by other documents cannot be detected by this type of analysis. Therefore, results from studies that analyse citations to identify relative differences in the sizes of bibliographic databases are likely to be affected by these confounding factors.

Of the six data sources that are analysed in this study, only two (Microsoft Academic and COCI) offered free and unrestricted access to the complete list of documents (or citation relationships in the case of COCI) that they covered at the time of data collection, although Dimensions now also offers this to researchers. To include all data sources in this study in a comparable way, the alternative method (selection of seed sample and analysis of citations) was used to discover relative coverage differences among data sources across subject categories. Since citation extraction discrepancies seem likely to be small compared to coverage differences, the results should also be useful to detect differences in coverage between sources.

### Selection of seed sample

The sample of citations analysed in this paper was taken from a seed sample of highly-cited documents: those listed in Google Scholar's *Classic Papers* product[10] (GSCP). This sample comprises the top 10 most cited documents published in 2006 according to Google Scholar in each of 252 subject categories (except *French Studies*, which has only 5 documents). The 252 subject categories are also assigned to one or more of 8 broad subject areas. The seed sample contains a total of 2515 highly-cited documents. For more information on GSCP, see Orduna-Malea et al. (2018).

This seed sample was considered useful for the purpose of this study, as it is the only sample of documents in Google Scholar for which an article-level subject classification is available. At the time of data collection, no other sample of documents with an article-level classification was readily available to us, and a sample with these characteristics was considered superior to the journal-level classification schemes that are used in sources such as Web of Science and Scopus. Additionally, being aware of the difficulties that extracting data from Google Scholar entail (Else 2018), the election of a sample of documents that were known to be highly cited also guaranteed a high efficiency in the citation extraction

[9] Visser et al. (2020) found that a large number of citations missing from Microsoft Academic were caused by missing reference lists in the citing documents. As far as we know no study has analysed how many missing citations in Google Scholar are caused by missing reference lists.

[10] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006.

process (each request to Google Scholar retrieved the maximum amount of records that the search engine displays per page).

This study analyses the complete list of documents that cite this seed sample, as reported in a variety of citation indexes (Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and COCI). In this study, they are called citing documents, or more simply, citations. Thus, this study follows the same approach as Martín-Martín et al. (2018).

### Collection of citation data

Each of the 2515 highly-cited documents were searched on Google Scholar, Microsoft Academic, Scopus, Dimensions, WoS, and COCI (Table 1). For each seed document found in a data source, the list of citing documents was extracted, as described below. The searches and data extraction were carried out in May and June 2019 (i.e., not re-using the data from the previous paper).

Google Scholar has no data exporting capabilities in its web interface and no API. Instead, a custom web scraper was used to extract the list of citing documents for each highly-cited document in the seed sample (Martín-Martín 2018). CAPTCHAs were solved manually when they appeared.

Google Scholar provides up to 1000 results per query. In order to download the complete list of citing documents for those with more than 1000 citations, queries were split by the publication year of the citing documents. Using this method, we were able to download most of the citing documents available in Google Scholar: for 2429 (96.5%) seed documents, we were able to extract a list of citing documents, amounting to at least 98% of the total citation counts reported by Google Scholar for these seed documents. In eight cases (extremely highly-cited seed documents), splitting queries by publication year was not enough to find all possible citing documents, and in these cases the number of citing documents extracted from Google Scholar was lower than 75% of the reported Google Scholar citation counts. This disadvantages Google Scholar in comparison to the other sources, for which all citing documents could be extracted. 2,689,809 citations were extracted from Google Scholar.

The metadata provided by Google Scholar is limited (Delgado López-Cózar et al. 2019). For example, Google Scholar does not provide the DOI of a document, which is very useful for document matching across data sources, and therefore relevant to our study. To enrich the limited metadata provided by Google Scholar, we followed several approaches. First, given that most of the citing documents from Google Scholar had already been analysed (Martín-Martín et al. 2018), we matched the newly extracted list of citing documents to the data from the previous study, and retrieved all the enriched metadata that was available in the dataset used for the 2018 study. Next, for all the citing documents that could not be matched in the previous step (mostly newer citations), metadata was extracted from the HTML Meta tags in the landing page of each citing document, and with public metadata APIs when a CrossRef or DataCite DOI could be found. These methods produced a DOI for 62.9% of all Google Scholar citations.

To collect citation data from Microsoft Academic, the Academic Search API[11] was used. This API is free with a limit of 10,000 transactions per month. At the moment of data collection, this API did not facilitate searching directly by DOI (Thelwall 2018b). For this

[11] https://msr-apis.portal.azure-api.net/docs/services/academic-search-api.

**Table 1** No. of seed highly-cited documents and citations found in each data source

| Source | Seed documents[a] | | Citations |
|---|---|---|---|
| | *N* | % | |
| Google Scholar | 2515 | 100 | 2,689,809 |
| Microsoft Academic | 2500 | 99.4 | 1840,702 |
| Scopus | 2447 | 97.3 | 1,738,573 |
| Dimensions | 2478 | 98.5 | 1,649,162 |
| WoS | 2342 | 93.1 | 1,503,657 |
| COCI | 2471 | 98.3 | 852,413 |

[a]Due to the sample selection process, the figures related to the seed documents found in each data source cannot be used as evidence that Google Scholar has higher coverage than the other sources

reason, every highly-cited seed document was first searched for by title. Once the seed document was retrieved and confirmed to be correct, new queries were submitted to retrieve the list of citing documents. Up to 1000 citing documents per query could be extracted (seed documents with over 1000 citations required more than one query to extract all citations). For each citing document, the Microsoft Academic internal Id, as well as the DOI, the document title, the list of authors, the publication year, the language, and the citation counts, were retrieved. 1,840,702 citations were extracted from Microsoft Academic.

To collect citation data from Scopus, the exporting capabilities of the web interface were used. Each seed highly-cited document was searched in Scopus by DOI and title, and, if found, the list of citing documents was exported in csv format. Scopus allows 2000 records per query to be exported. When seed documents had over 2000 citations, the alternative email service was used, which allows 20,000 records to be exported. No document in the seed sample had more than 20,000 citations in Scopus. 1,738,573 citations were extracted from Scopus.

To collect citation data from Dimensions, its API was used, which is free for research.[12] The Dimensions API allows searching by DOI. Therefore, all seed highly-cited documents were searched for using their DOI, and, when unavailable, by their title. Once all the seed documents had been identified in Dimensions, the API was also used to extract the list of citing documents. For each citation, the basic bibliographic information (DOI, title, authors, publication year, source, document type) was recorded. 1,649,162 citations were extracted from Dimensions.

To collect citation data from WoS, the web interface was used. All citation indexes in WoS Core Collection were included in the analysis, including the Emerging Sources Citation Index (from publication year 2005 to the present). Each seed highly-cited document was searched by its DOI, and, when unavailable, by its title. The list of citing documents was then exported in batches of up to 500. The exported files were consolidated into a single table using a set of R functions (Martín-Martín and Delgado López-Cózar 2016). 1,503,657 citations were extracted from WoS.

To collect citation data from COCI, the public API was used. The DOI of each seed highly-cited document was searched in order to retrieve the complete list of citing DOIs. 852,413 citation relationships were extracted from COCI.

[12] https://www.dimensions.ai/scientometric-research/.

### Analysis of citation data

To calculate citation overlaps across data sources, the citing documents from different data sources were matched. The matching process started with two data sources (WoS and Scopus), and the result was a full join of the two sources: a table containing all citations found both by WoS and Scopus, as well as the citations found only by one of the data sources. The resulting dataset was matched to the data obtained from another data source (Dimensions), and this process was repeated until all data sources were merged into a master list of citations (Table 2). The matching criteria are below, and are the same as previously used (Martín-Martín et al. 2018):

1. For each pair of data sources *A* and *B* and a seed highly-cited document *X*, all citing documents with a DOI that cite *X* according to *A* where matched to all citing documents with a DOI that cite *X* according to *B*.
2. For each of the unmatched documents citing *X* in *A* and *B*, a further comparison was carried out (except in the matching round where COCI data was integrated into the master table). The title of each unmatched document citing *X* in *A* was compared to the titles of all the unmatched documents citing *X* in *B*, using the restricted Damerau-Levenshtein distance (optimal string alignment) (Damerau 1964; Levenshtein 1966). The pair of citing documents which returned the highest title similarity (1 is perfect similarity) was selected as a potential match. This match was considered successful if either of the following conservative heuristics was met:
    - The title similarity was at least 0.8, and the title of the citing document was at least 30 characters long (to avoid matches between short, undescriptive titles such as "Introduction").
    - The title similarity was at least 0.7, and the first author of the citing document was the same in *A* and *B*.

The matching criteria described above are intentionally conservative, so a match is only accepted when the two documents have very similar metadata. The analysis does not attempt to remove duplicate citations within the same data source, although Google Scholar and Scopus (and perhaps others) are afflicted by this issue (Orduna-Malea et al. 2017; van Eck and Waltman 2019). In this study, if there are duplicate citations within the same data source only one of the instances will be linked to the same citation in other sources, while the rest will (erroneously) appear as unique citations. Therefore, the percentage overlaps between sources calculated are conservative estimates (i.e., they might be higher than reported here). A replication of the overlap analysis carried in Martín-Martín et al. (2018) for one subject category (Operations Management) showed that overlap figures are affected little when duplicates are identified and removed, however (Chapman and Ellinger 2019).

Given that the objective is to detect relative differences in coverage across databases, to make comparisons as fair as possible the subset of citations that are considered in each comparison is adapted to include only citation relationships where the cited seed document is covered by all sources present in the comparison. For example, in a comparison of coverage across the six data sources analysed in this study (Table 1, top), only citations

**Table 2** Rounds of the matching process

| Matching round | Data sources being matched | Resulting dataset | Merged citations |
|---|---|---|---|
| 1st | WoS ⋈ Scopus | master_1 | 1,852,681 |
| 2nd | master_1 ⋈ Dimensions | master_2 | 1,990,862 |
| 3rd | master_2 ⋈ Microsoft Academic | master_3 | 2,263,896 |
| 4th | master_3 ⋈ COCI | master_4 | 2,273,067 |
| 5th | master_4 ⋈ Google Scholar | master_5 | 3,073,351 |

to the 2319 seed highly-cited documents covered by all six data sources are considered. However, in pairwise comparisons, such as the Venn diagram that represents overlapping and unique citations in Google Scholar and Microsoft Academic (Fig. 2a), the citations to the 2500 seed highly-cited documents that are known to be covered by these two sources were analysed.

Data processing was carried out with the R programming language (R Core Team 2014) using several R packages and custom functions (Dowle et al. 2018; Krassowski 2020; Larsson et al. 2018; Martín-Martín and Delgado López-Cózar 2016; van der Loo et al. 2018; Walker and Braglia 2018; Wickham 2016; Wilke 2019). The resulting data files are openly available.[13]

## Results

### Overall results (all subject categories)

#### Relative overlap

Overall, Google Scholar has the highest coverage, as it found 88% of all possible citations (2,918,105) to the 2319 highly-cited documents in our sample that were covered by the six sources under analysis (Fig. 1, first row). Microsoft Academic, Scopus, Dimensions and WoS found substantially fewer (60–52% of all citations). COCI found only 28% of all possible citations.

In terms of relative overlaps between two data sources, larger data sources are able to find a vast majority of the citations found by the smaller data sources (Fig. 1, row 2 through 6). Thus, Google Scholar found 89% of the citations in the second data source with the largest coverage (Microsoft Academic), and up to 94% of the citations in the smaller sources (WoS, COCI). On the other side of the spectrum, COCI, the smallest source, found between 30% and 51% of the citations found by the other sources (Google Scholar and Dimensions, respectively).

For Microsoft Academic, Scopus, Dimensions, and WoS, the relative overlap between any two of these sources ranges from high (WoS found 73% of the citations

[13] https://osf.io/gnb72/ (2019 folder).

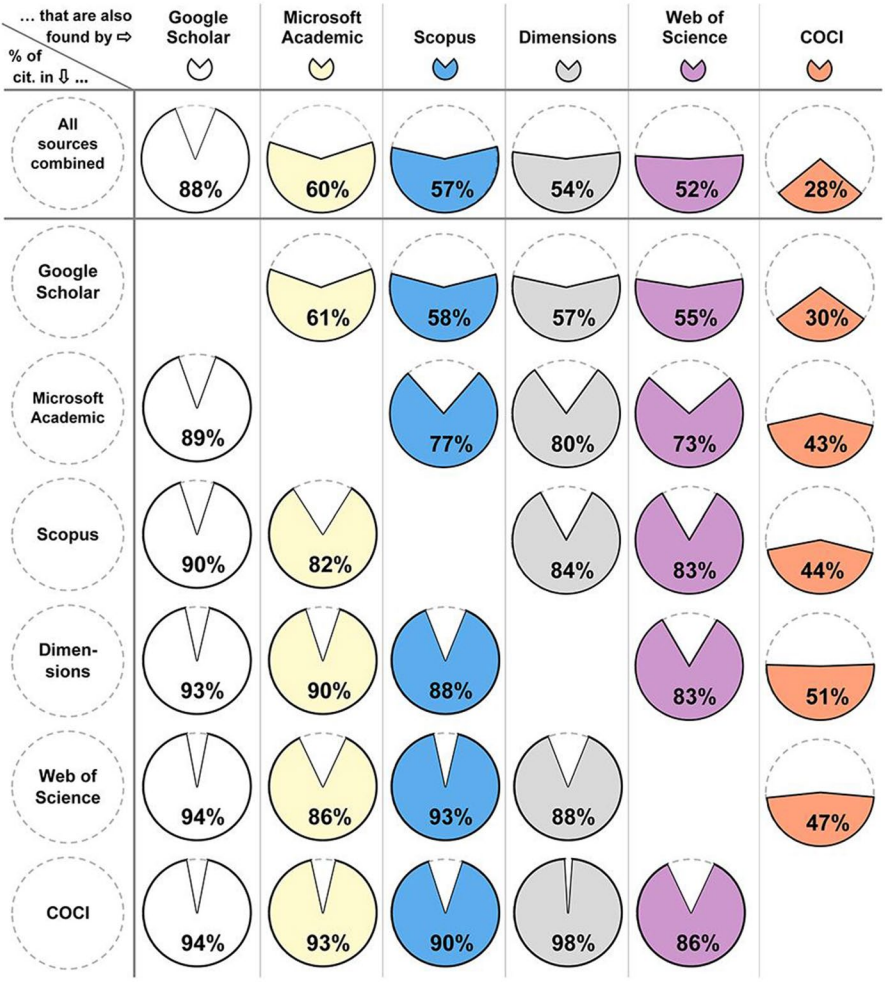


**Fig. 1** Percentage of citations found by each database, relative to all citations (first row), and relative to citations found by the other databases (subsequent rows)

found by Microsoft Academic) to almost full overlap (Dimensions found 98% of the citations available in COCI). Figure 1 shows that it is not always the case that the larger the source, the higher the proportion of citations from another source that it will be able to find. For example, Dimensions found 80% of the citations available in Microsoft Academic, while Scopus (larger than Dimensions) found 77%. The cause of this might be that both Microsoft Academic and Dimensions cover non-journal content, such as

preprints, while Scopus does not. Scopus found 93% of the citations found by WoS, while Microsoft Academic (larger than Scopus) found 86%. Dimensions was able to find the highest proportion of COCI citations (98%) out of all the other sources (including Google Scholar).

## Overlaps within the full set of citations

A quarter (26%) of all citations were found only by Google Scholar (Fig. 2), 21% of the citations were found by the six sources, while 18% were found by all sources except COCI. The remaining 35% were found by combinations of four or less data sources, and the highest values were found in sectors that include Google Scholar and/or Microsoft Academic.

When considering all possible pairwise combinations (Fig. 3), the pairs of data sources that are most similar in terms of full citation overlap are Scopus/WoS (78% of all citations found by either were found by both), followed by Scopus/Dimensions (75%), Dimensions/WoS (75%), and Microsoft Academic/Dimensions (74%). Pairs that include Google Scholar or COCI display lower percentages of overlap: in the case of Google Scholar this is caused by the extra coverage in Google Scholar that is not found in the other sources, while for COCI the reason is the opposite.

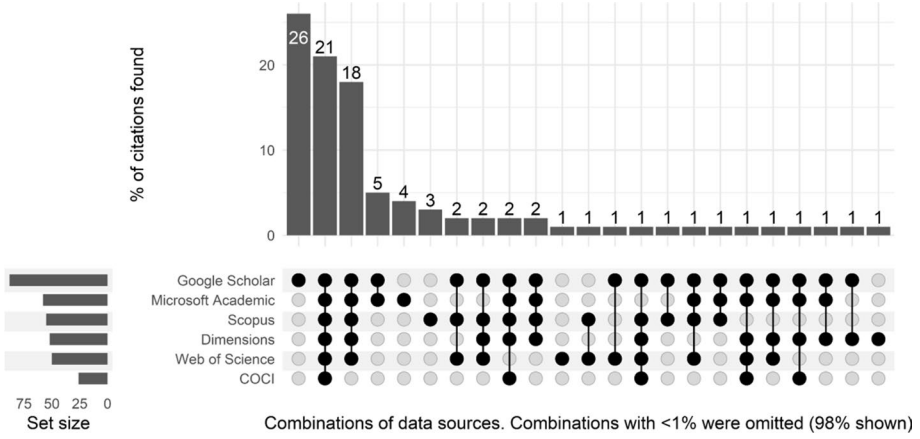


**Fig. 2** Overlaps of citations found by Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and COCI. Values expressed as percentages relative to $N = 2{,}918{,}105$ citations to 2319 documents. Combinations with values below 1% are not displayed

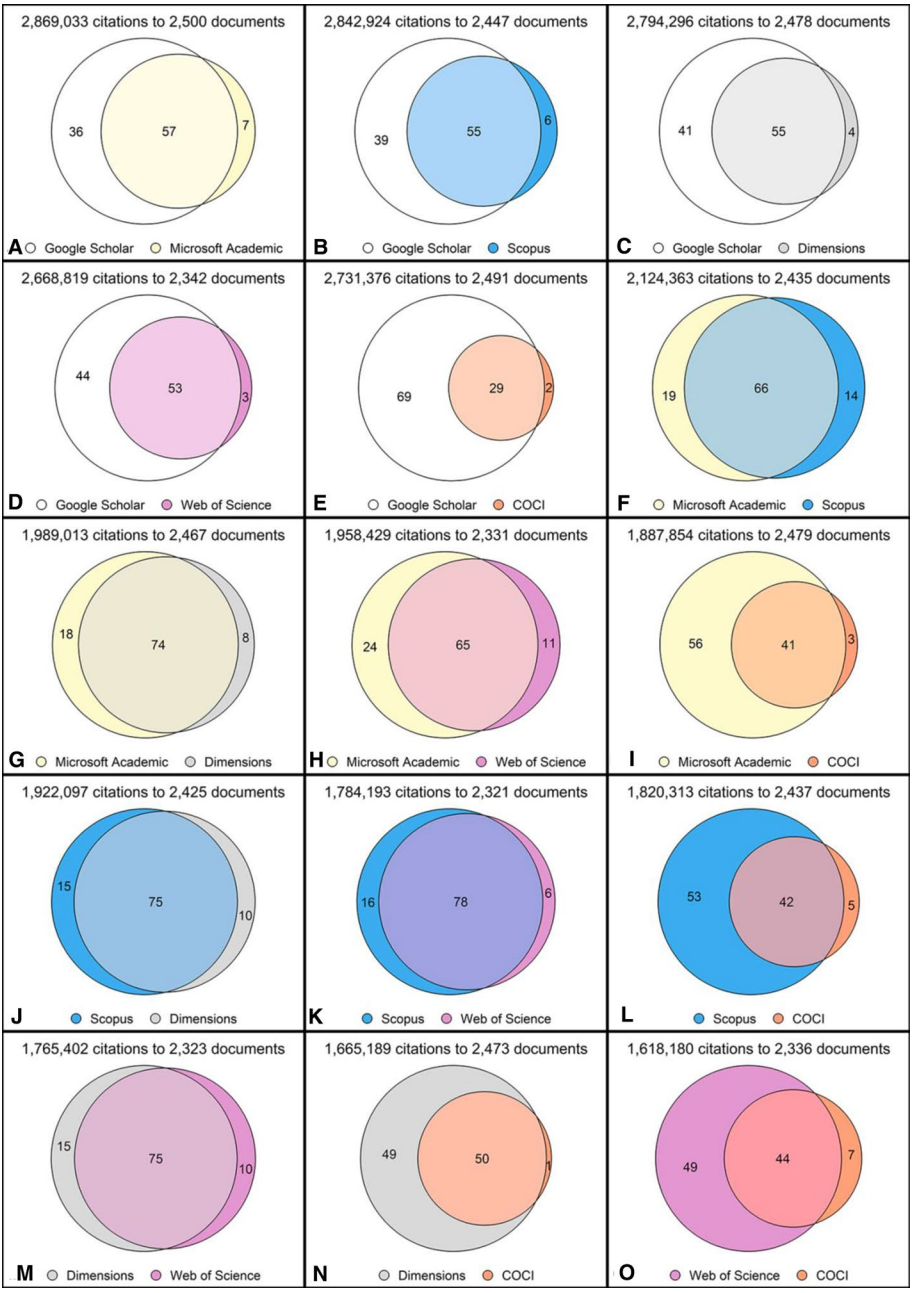


**Fig. 3** Comparison of citing document overlaps between Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and COCI (pairwise). Figures within Venn diagrams expressed as percentages

## Analysis by subject areas and categories

### Relative overlap

Disaggregating the data by broad subject areas provides a more detailed picture of the coverage of these sources. Google Scholar found the great majority of citations (85–90%) in all eight subject areas (Table 3) and COCI found the fewest. COCI has differences in coverage across areas: in the Humanities and Social Sciences it found 18–20% of all citations, while in the STEM areas (Science, Technology, Engineering, and Mathematics) it found a higher proportion of citations (27–32%).

Between these two extremes, the other four sources (Microsoft Academic, Scopus, Dimensions, and WoS) tend to have similar coverage of each field, but differences between fields (Table 3). They have more comprehensive coverage for *Chemical & Material Sciences* (69–72%), followed by *Life Sciences & Earth Sciences* (60–68%). Conversely, their coverage is much lower in *Humanities, Literature & Arts* (25–39%), *Social Sciences* (33–47%) and *Business, Economics & Management* (29–47%). Among these four, Microsoft Academic seems to have the most comprehensive coverage, except in *Physics & Mathematics*, where it found fewer of the citations (57%) than the other sources.

Further disaggregating the data to identify the percentage of relative citation overlap for each pair of sources in each subject area (Table 4), the patterns for the complete dataset (Fig. 1) recur. Google Scholar consistently found most citations found by the other sources across all areas; there is a higher relative overlap between Microsoft Academic and Dimensions/COCI than between Microsoft Academic and Scopus/WoS; conversely, the relative overlap between Scopus and WoS is always higher than between Scopus and other sources; the highest relative overlap in each area is always for Dimensions/COCI; Microsoft Academic seems to lack coverage in *Physics & Mathematics*, as evidenced by its lower relative overlap in this area.

**Table 3** Percentage of citations found by each data source, relative to the total number of citations found overall and by broad areas

| | N | % of citations found (relative to N) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
| Humanities, Literature & Arts | 89,337 | 87% | 39% | 31% | 29% | 25% | 18% |
| Social Sciences | 406,661 | 88% | 47% | 40% | 36% | 33% | 20% |
| Business, Economics & Management | 235,338 | 88% | 47% | 34% | 32% | 29% | 19% |
| Engineering & Computer Science | 691,164 | 88% | 63% | 61% | 54% | 48% | 30% |
| Physics & Mathematics | 317,320 | 90% | 57% | 64% | 59% | 59% | 36% |
| Health & Medical Sciences | 1,001,507 | 85% | 63% | 59% | 58% | 51% | 27% |
| Life Sciences & Earth Sciences | 571,817 | 89% | 68% | 64% | 63% | 60% | 32% |
| Chemical & Material Sciences | 253,990 | 90% | 69% | 75% | 72% | 72% | 32% |

**Table 4** Relative pairwise overlaps between data sources (%). Overall and by broad subject areas

**A Humanities, Literature & Arts**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 39% | 33% | 30% | 29% | 19% |
| Microsoft Acad. | 86% | | 57% | 62% | 53% | 42% |
| Scopus | 84% | 68% | | 65% | 68% | 42% |
| Dimensions | 89% | 86% | 75% | | 69% | 59% |
| Web of Science | 87% | 73% | 80% | 70% | | 46% |
| COCI | 93% | 92% | 77% | 94% | 73% | |

**B Social Sciences**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 48% | 41% | 39% | 37% | 22% |
| Microsoft Acad. | 88% | | 66% | 69% | 60% | 40% |
| Scopus | 89% | 78% | | 75% | 76% | 43% |
| Dimensions | 93% | 90% | 83% | | 76% | 54% |
| Web of Science | 92% | 82% | 88% | 81% | | 47% |
| COCI | 96% | 95% | 85% | 96% | 80% | |

**C Business, Economics & Management**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 46% | 35% | 34% | 31% | 20% |
| Microsoft Acad. | 85% | | 58% | 61% | 52% | 36% |
| Scopus | 91% | 80% | | 77% | 75% | 45% |
| Dimensions | 93% | 90% | 82% | | 75% | 55% |
| Web of Science | 93% | 84% | 87% | 83% | | 50% |
| COCI | 94% | 92% | 83% | 95% | 78% | |

**D Engineering & Computer Science**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 65% | 62% | 58% | 55% | 32% |
| Microsoft Acad. | 90% | | 79% | 78% | 70% | 43% |
| Scopus | 89% | 82% | | 81% | 79% | 45% |
| Dimensions | 93% | 91% | 91% | | 82% | 53% |
| Web of Science | 93% | 86% | 94% | 87% | | 49% |
| COCI | 94% | 94% | 92% | 97% | 83% | |

**E Physics & Mathematics**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 58% | 65% | 61% | 61% | 37% |
| Microsoft Acad. | 91% | | 83% | 83% | 78% | 48% |
| Scopus | 91% | 74% | | 85% | 87% | 52% |
| Dimensions | 93% | 80% | 93% | | 88% | 60% |
| Web of Science | 93% | 75% | 95% | 88% | | 55% |
| COCI | 92% | 77% | 94% | 98% | 90% | |

**Table 4** (continued)

**F Health & Medical Sciences**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 64% | 61% | 62% | 58% | 29% |
| Microsoft Acad. | 87% | | 78% | 84% | 75% | 41% |
| Scopus | 88% | 84% | | 86% | 84% | 40% |
| Dimensions | 91% | 91% | 86% | | 82% | 45% |
| Web of Science | 95% | 87% | 92% | 89% | | 43% |
| COCI | 94% | 96% | 89% | 99% | 86% | |

**G Life Sciences & Earth Sciences**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 69% | 67% | 67% | 64% | 34% |
| Microsoft Acad. | 91% | | 82% | 86% | 80% | 45% |
| Scopus | 93% | 88% | | 88% | 88% | 46% |
| Dimensions | 94% | 93% | 90% | | 87% | 50% |
| Web of Science | 95% | 91% | 94% | 91% | | 48% |
| COCI | 96% | 96% | 92% | 98% | 90% | |

**H Chemical & Material Sciences**

| Percentage of citations in ⇩ ... / ... that are also found by ⇨ | Google Scholar | Microsoft Academic | Scopus | Dimensions | Web of Science | COCI |
|---|---|---|---|---|---|---|
| Google Scholar | | 71% | 78% | 75% | 75% | 34% |
| Microsoft Acad. | 93% | | 90% | 92% | 88% | 43% |
| Scopus | 93% | 83% | | 89% | 92% | 40% |
| Dimensions | 94% | 89% | 94% | | 91% | 44% |
| Web of Science | 94% | 84% | 96% | 90% | | 41% |
| COCI | 95% | 93% | 93% | 98% | 91% | |

## Full overlap

The differences in coverage between the older (Google Scholar, Scopus, WoS) and newer (Microsoft Academic, Dimensions) sources across subject areas are also evident from three-way comparisons (Figs. 4, 6, 8). The three-set combinations of data sources that are not displayed here are accessible from "Appendix 1". The combinations that include more than one of the older sources are not included here because they were discussed in a previous study (Martín-Martín et al. 2018) and the results have barely changed. The combinations that include COCI are not displayed here because it is essentially a subset of the other sources (especially Dimensions).

Venn diagrams for the 252 specific subject categories are also accessible from "Appendix 1". Figures 5, 7, 9 and 10 display the distribution of the proportions of citations that would fall in each section of the Venn diagrams calculated at this level of aggregation, for various pairs of data sources. The remaining combinations are accessible from "Appendix 2".

## Google Scholar and the new sources: Microsoft Academic, and Dimensions

For Google Scholar, Microsoft Academic, and Dimensions, the largest percentages of full overlap (citations found by the three sources) occur in the STEM fields (Fig. 4). These

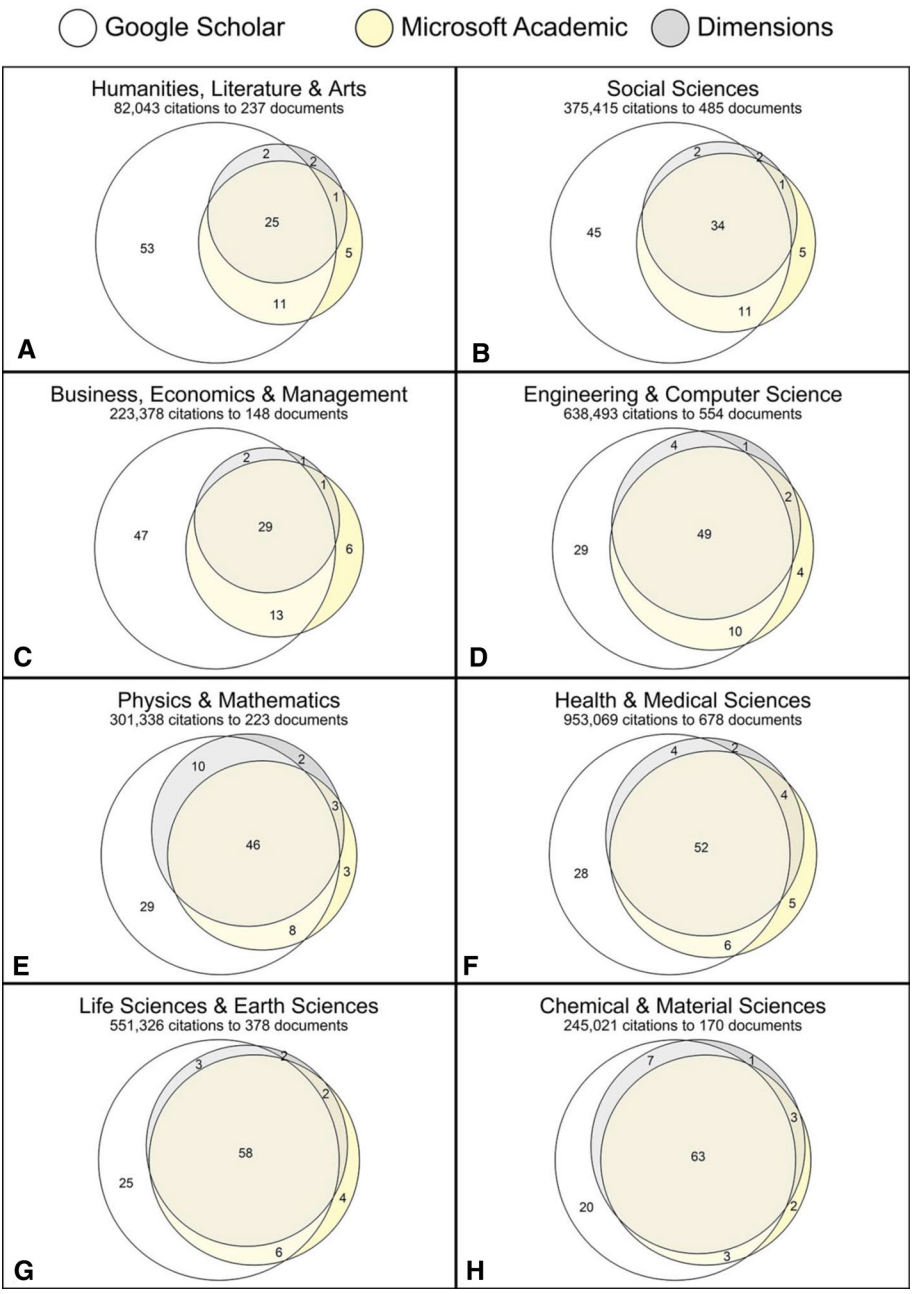


**Fig. 4** Overlaps between citations found by Google Scholar, Microsoft Academic, and Dimensions in broad subject areas. Figures within Venn diagrams expressed as percentages

range from 46% in *Physics and Mathematics*, to 63% in *Chemical and Material Sciences.* Full overlap in the areas of Humanities and Social Sciences is distinctly lower (25–34%). This is caused by lower coverage of these areas in Microsoft Academic and Dimensions.


**Fig. 5** Distribution of citations that fall within each sector of the Venn diagrams that compare Google Scholar to Microsoft Academic and Dimensions. Calculated at the level of subject categories, and aggregated by subject areas

The percentage of citations in Microsoft Academic and/or Dimensions that is not covered by Google Scholar ranges from 6% (in *Chemical and Material Sciences*) to 11% (in *Health & Medical Sciences*).

At the level of specific subject categories, for pairwise comparisons between Google Scholar/Microsoft Academic and Google Scholar/Dimensions (Fig. 5) the general trend of the subject area is followed, with variations in some subject categories. The variation seems to be higher between Google Scholar/Dimensions than between Google Scholar/Microsoft Academic. Nevertheless, in both comparisons the percentages in the sector "Only in Google Scholar" are higher in the Humanities and Social Sciences, and lower in STEM fields. The sector "In both data sources" almost mirrors the one above, and the sectors "Only in Microsoft Academic" and "Only in Dimensions" have values almost exclusively below 10%, with two major exceptions. These correspond to the categories *Astronomy & Astrophysics,*[14] and *Psychology.*[15] In these two categories, many citations found by Microsoft Academic and Dimensions were not found by Google Scholar. In the case of Psychology, the low citation coverage in Google Scholar is caused by one extremely highly-cited document (*Using thematic analysis in psychology*, by Virginia Braun and Victoria Clarke[16]), which at the time of data collection had 54,323 citations in Google Scholar. However, because of the limitations of Google Scholar's search interface for data extraction, only 10,996 citations could be extracted from Google Scholar for this article.

### Scopus and the new sources: Microsoft Academic and Dimensions

For Microsoft Academic, Scopus, and Dimensions, none of the sources is always larger than the others, with the results varying by subject area (Fig. 6). Microsoft Academic sometimes has larger coverage than Scopus (Humanities and Social Sciences), although in these areas both contribute many unique citations. Scopus also sometimes provides more coverage than Microsoft Academic (*Physics & Mathematics*, *Chemical & Material Sciences*). The previously seen trend of higher percentages of full overlap in STEM fields also occurs here. The number of citations found by Dimensions is similar to that of Scopus across all subject areas, but there are also many citations that one of them finds that the other does not in the Humanities and Social Sciences. Comparing the three sources together, Dimensions provides the fewest unique citations.

In most subject categories (Fig. 7), there are large Microsoft Academic/Scopus and Scopus/Dimensions citation overlaps. This is especially evident in STEM categories, where

[14] Google Scholar/Microsoft Academic: https://osf.io/g8z42/; Google Scholar/Dimensions: https://osf.io/bwv5s/.

[15] Google Scholar/Microsoft Academic: https://osf.io/jqwah/; Google Scholar/Dimensions: https://osf.io/xnf24/.

[16] https://www.tandfonline.com/doi/abs/10.1191/1478088706QP063OA.

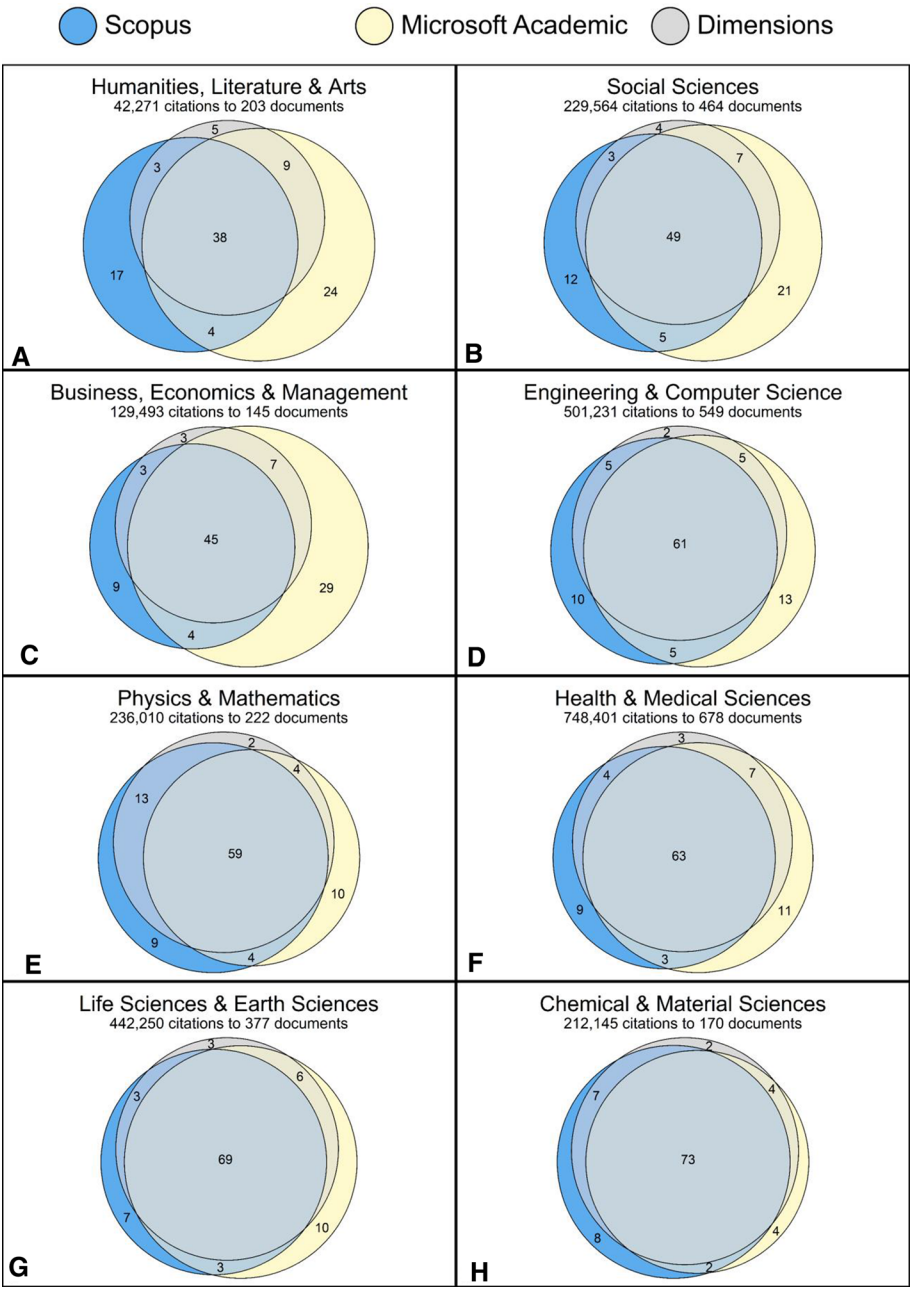


**Fig. 6** Overlap between citations found by Scopus, Microsoft Academic, and Dimensions, by broad subject area. Figures within Venn diagrams expressed as percentages

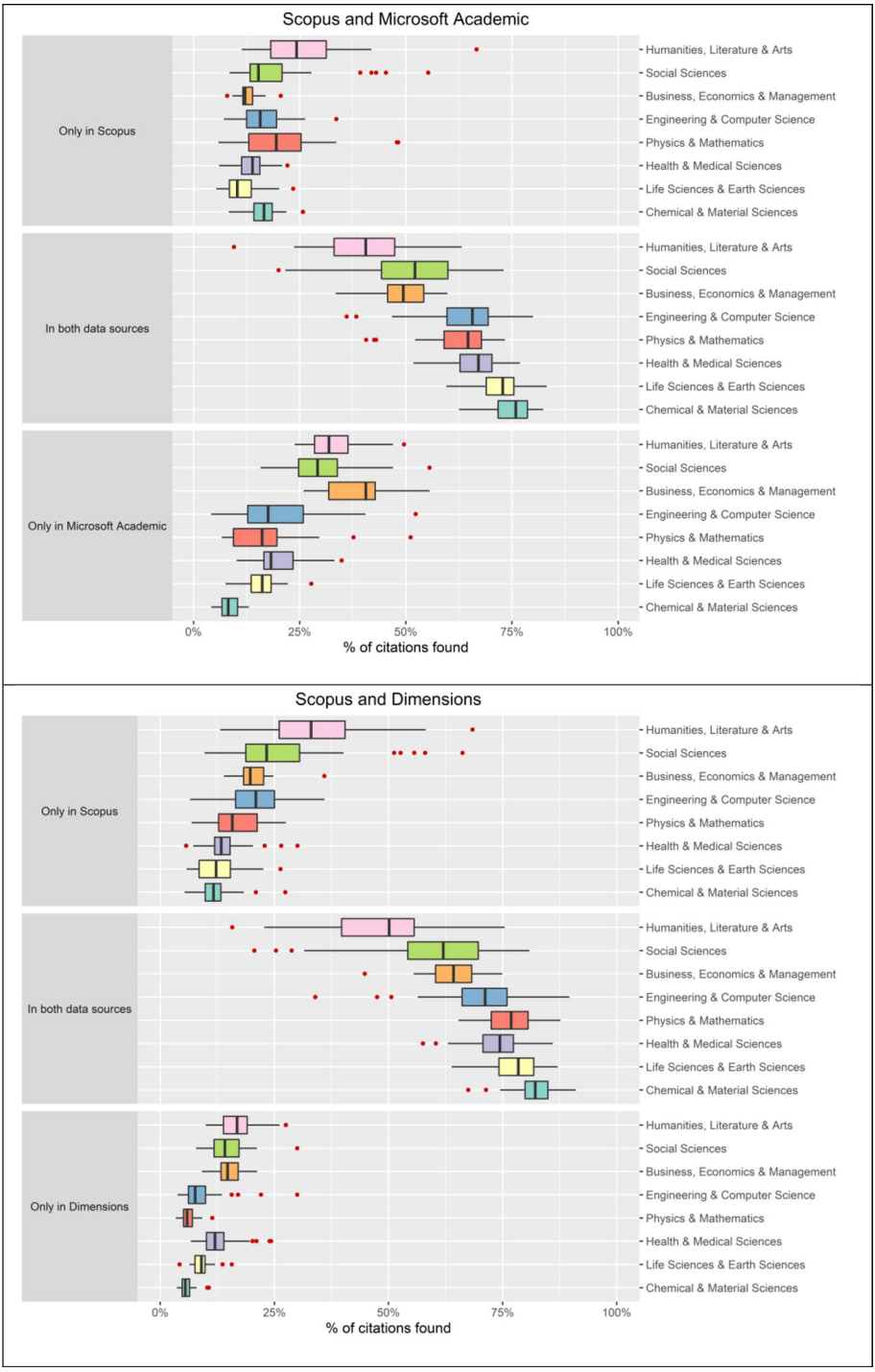


**Fig. 7** Distribution of citations within each sector of the Venn diagrams that compare Scopus to Microsoft Academic and Dimensions. Calculated at the level of subject categories, and aggregated by subject areas

the overlap in almost all cases exceeds 50%. For Microsoft Academic/Scopus (Fig. 7, top), there are 66 (out of 252) subject categories where the overlap exceeds 70%, and for Scopus/Dimensions, 148 categories exceed this overlap. Extreme cases of low overlap between sources are almost always in the Humanities and Social Sciences. For Microsoft Academic/Scopus, the lowest overlaps (below 30%) are in French Studies[17] (9%, although in this case the results are based only on citations to one seed document, because the rest were not covered by Microsoft Academic and Scopus), International Law[18] (20%), European Law[19] (21%), American Literature & Studies[20] (24%), Law[21] (26%), and Film[22] (27%). In 182 categories (out of 252) Microsoft Academic found more citations than Scopus. There are also some outlier cases of low overlap in STEM categories, such as over 50% of citations in Computer Graphics[23] and Discrete Mathematics[24] only being available in Microsoft Academic (compared to Scopus), or 48% of citations in High Energy & Nuclear Physics[25] and Quantum Mechanics[26] only being found by Scopus (compared to Microsoft Academic). For Scopus/Dimensions (Fig. 7, bottom), many of the same categories have the lowest overlap: French Studies,[27] International Law,[28] American Literature & Studies,[29] European Law,[30] and History.[31] These low coverage figures are caused by Microsoft Academic and Dimensions having a lower coverage of citations in these categories than Scopus. In 36 categories (out of 252) Dimensions found more citations than Scopus.

### Web of Science and the new sources: Microsoft Academic and Dimensions

Comparing Microsoft Academic, Dimensions and WoS (Fig. 8), there are many unique citations in Microsoft Academic and WoS. Out of these three, Dimensions found the fewest unique citations (2–6% depending on the area). Again, the divergence is higher in the Humanities and Social Sciences, where Microsoft Academic has the highest percentages of unique citations. Microsoft Academic also has lower coverage in *Physics & Mathematics* and (to a lower degree) in *Chemical & Material Sciences*.

The results by subject category confirm that some categories deviate from the trend in a broad area (Fig. 9). Considering Microsoft Academic/WoS (Fig. 9, top), Microsoft Academic's coverage is large compared to WoS for *Computing Systems*[32] (73% of all citations),

[17] https://osf.io/gmrju/.
[18] https://osf.io/bzha2/.
[19] https://osf.io/f36sn/.
[20] https://osf.io/7qzmk/.
[21] https://osf.io/4gtdc/.
[22] https://osf.io/ctzb7/.
[23] https://osf.io/rz4cj/.
[24] https://osf.io/v6bgy/.
[25] https://osf.io/vafzp/.
[26] https://osf.io/87cdh/.
[27] https://osf.io/bqpz4/.
[28] https://osf.io/p26ua/.
[29] https://osf.io/fngph/.
[30] https://osf.io/pdnxt/.
[31] https://osf.io/xjhfw/.
[32] https://osf.io/ugvh3/.

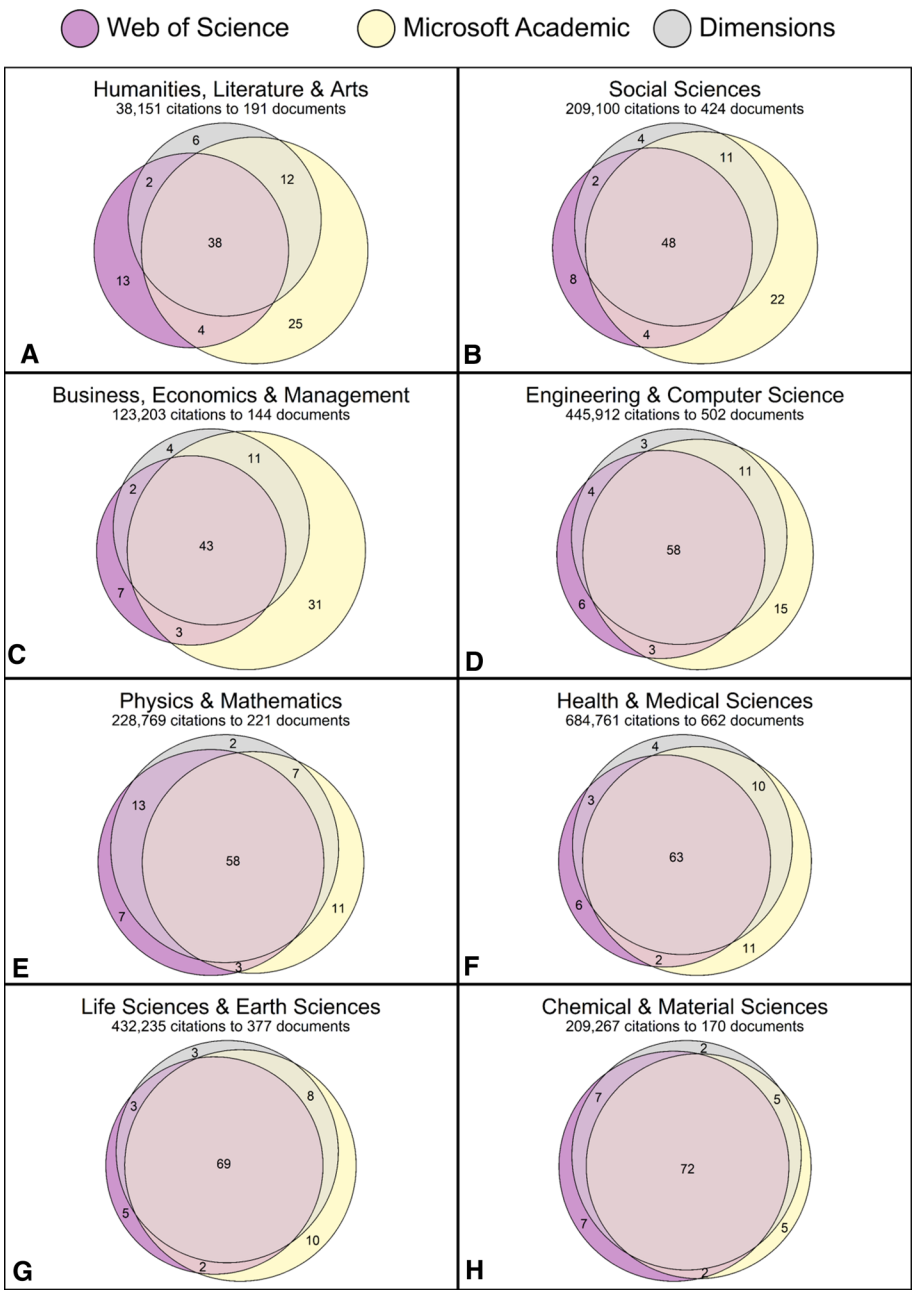


**Fig. 8** Overlaps between citations found by Web of Science, Microsoft Academic, and Dimensions, by broad subject areas. Figures within Venn diagrams expressed as percentages

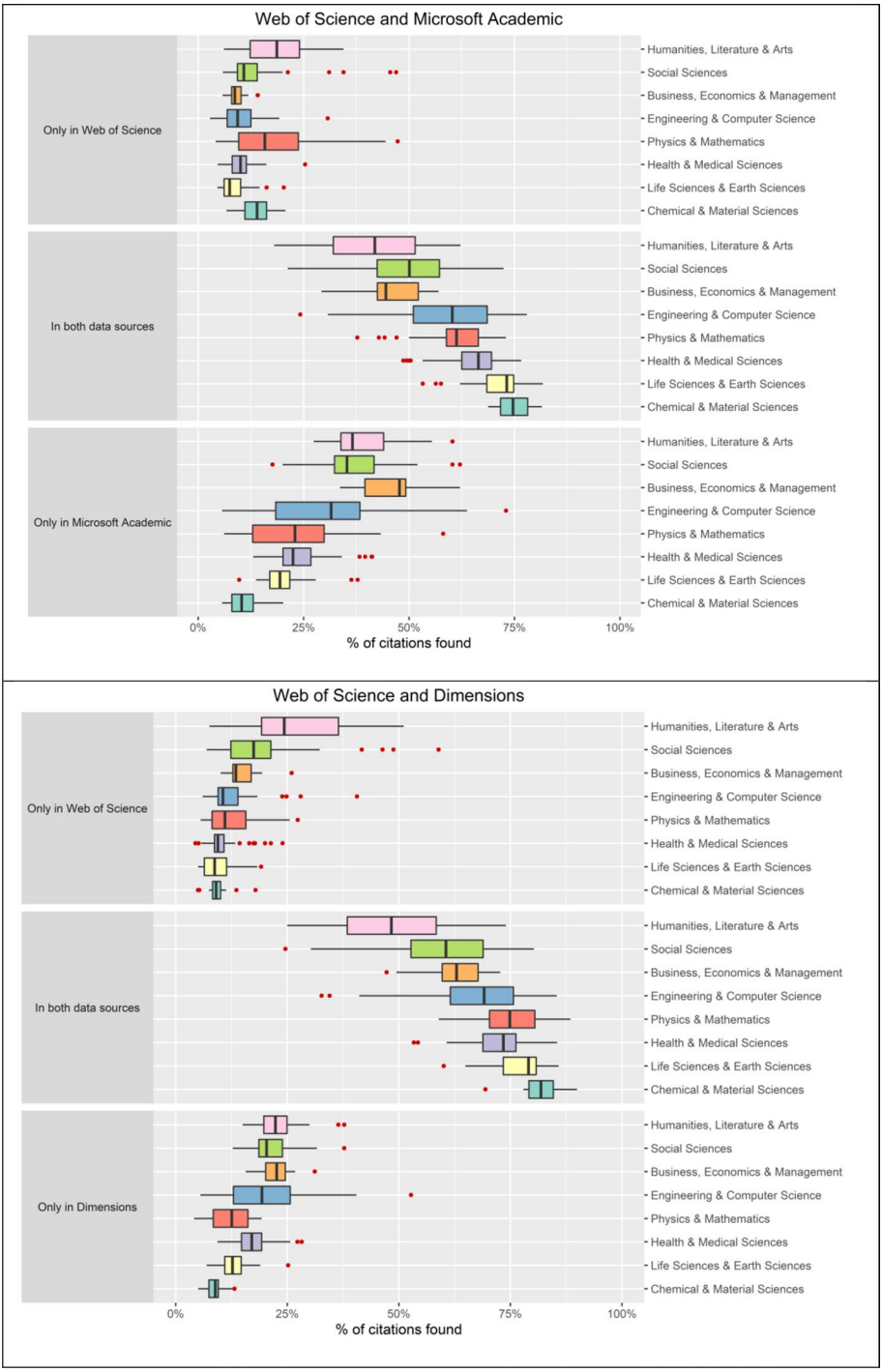


**Fig. 9** Distribution of citations within each sector of the Venn diagrams that compare Web of Science to Microsoft Academic and Dimensions. Calculated at the level of subject categories, and aggregated by subject areas

*Software* Systems[33] (63%), *Educational Administration*[34] (62%), Chinese *Studies & History*[35] (60%), and *Discrete Mathematics*[36] (58%). The gaps in coverage of Microsoft Academic in *International Law,*[37] and *Law*[38] occur again here, as 47% and 46% of the citations in these categories (respectively) are only found by WoS. Something similar occurs in the categories included in *Physics & Mathematics:* the distribution of citations only found by WoS in this area has an unusually wide interquartile range when compared with the other areas, which is a sign that Microsoft Academic's gaps in coverage in this area affect more than one category. The most extreme cases are again *Quantum Mechanics*[39] and *High Energy & Nuclear Physics,*[40] with 47% and 44% of citations only found by WoS (respectively). In 223 categories (out of 252) Microsoft Academic found more citations than WoS. For the distributions of overlap and unique citations between Dimensions/WoS (Fig. 9, bottom), there are some similarities with the previous comparison: 51% of the citations in *Computing* Systems[41] are only found by Dimensions, and in Humanities and Social Sciences over a third of the citations in *Chinese Studies & History,*[42] and *Foreign Language Learning*[43] are only found by Dimensions, which reveals coverage gaps in these categories in WoS. In other Humanities categories, such as *American Literature & Studies*[44] (51%), *History*[45] (46%), or *Literature & Writing*[46] (46%) WoS found more unique citations than Dimensions. Dimensions also has gaps in coverage in *Computer Graphics,*[47] *International Law,*[48] *Law,*[49] and *Middle Eastern & Islamic Studies,*[50] compared to WoS. In 185 categories (out of 252) Dimensions found more citations than WoS.

### Microsoft Academic and Dimensions

At the level of subject categories, the vast majority of citations in Microsoft Academic/Dimensions are found either by both databases, or only by Microsoft Academic. In 209 out of 252 categories, the percentage of unique citations in Dimensions is below 10% (Fig. 10). The exceptions are in *Physics & Mathematics*, where 45% of the citations in *Quantum Mechanics,*[51] 39% of the citations in *High Energy & Nuclear Physics,*[52] and 26% of the

[33] https://osf.io/6vrnp/.
[34] https://osf.io/x9g3e/.
[35] https://osf.io/54xky/.
[36] https://osf.io/fa8sr/.
[37] https://osf.io/9584j/.
[38] https://osf.io/h7jt2/.
[39] https://osf.io/ghws2/.
[40] https://osf.io/gpyse/.
[41] https://osf.io/rsj4m/.
[42] https://osf.io/bvr3p/.
[43] https://osf.io/vmdbx/.
[44] https://osf.io/zd53e/.
[45] https://osf.io/q529p/.
[46] https://osf.io/qcdsh/.
[47] https://osf.io/sfd2g/.
[48] https://osf.io/a9mtx/.
[49] https://osf.io/n2e98/.
[50] https://osf.io/za5ks/.
[51] https://osf.io/3npwu/.
[52] https://osf.io/7qb8v/.

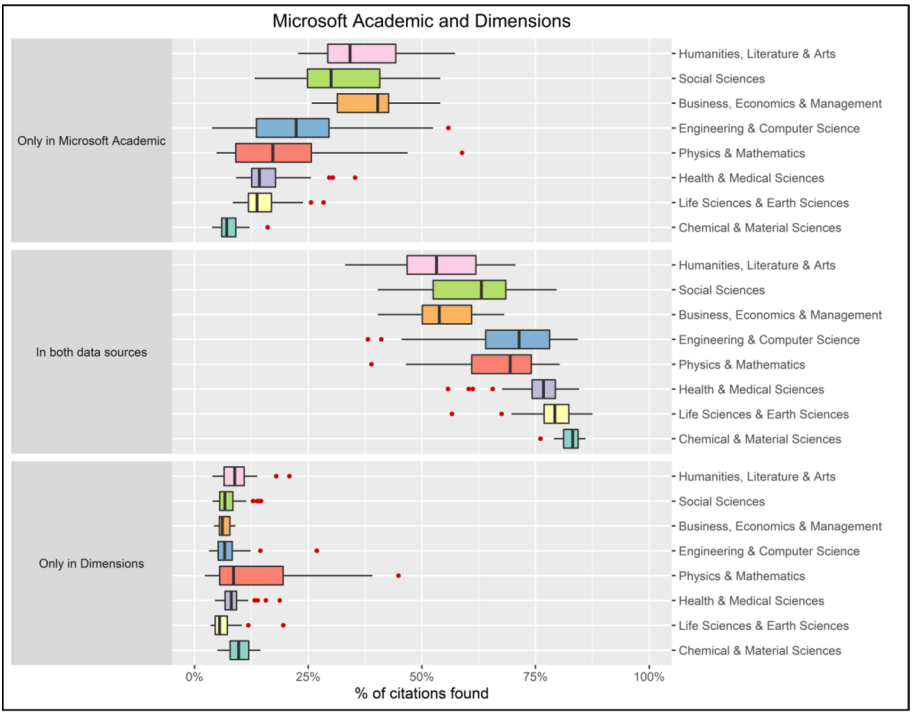


**Fig. 10** Distribution of citations within each sector of the Venn diagrams that compare Microsoft Academic and Dimensions. Calculated at the level of subject categories, and aggregated by subject areas

citations in *Plasma & Fusion*[53] (also included in *Engineering & Computer Science*) are only found by Dimensions. This again reveals the gap in coverage of Microsoft Academic in these categories. In 226 categories (out of 252) Microsoft Academic found more citations than Dimensions.

## Discussion

### Limitations

Because this study uses an updated and extended version of the sample used in Martín-Martín et al. (2018), many of the limitations declared in that study are also applicable here, as summarised below.

- The seed sample of documents used all highly-cited documents published in English in 2006 according to Google Scholar's *Classic Papers* product. To generalize the results presented here, it must be assumed that the population of documents that cite these

[53] https://osf.io/n5j8v/.

highly-cited documents is comparable to the general population of citing documents within each subject category. This might not be true in some cases, because different topics within the same category might have different citation patterns (certain highly-cited topics within a category might be overrepresented). Also, these results probably do not represent the reality of coverage of academic literature published in languages other than English and literature about regionally-relevant topics, where Google Scholar, Microsoft Academic, Dimensions and COCI may have an advantage.

- Google Scholar might have an unfair advantage in this analysis, as the initial seed sample was selected from a list of the highest-cited documents in this source (the accuracy of citation detection of Google Scholar in this specific sample could be higher than the average accuracy of citation detection across all documents in Google Scholar, which is unknown). However, the correlations between the citation counts of citing documents available in Martín-Martín et al. (2018) suggest that this advantage is not substantial: when analysing documents from the entire distribution of citation counts (not only highly-cited documents), Google Scholar still consistently reports higher citation counts than WoS and Scopus, while providing essentially the same citation rankings at the document level in most subject categories as the other two sources.
- The algorithm that matches citations across data sources is intentionally conservative: it is set to minimize false positives, potentially at the expense of false negatives. Therefore, the percentages of overlap shown in this study are lower bounds.
- Unlike Martín-Martín et al. (2018), where citations from documents included in the ESCI Backfile for documents published between 2005 and 2014 were not included in the analysis, in this study all available citation data available in the citation indexes that are part of WoS Core Collection is analysed.
- Data extraction for this analysis was carried out in May/June 2019. However, the rapid development of these platforms may render the results obsolete in the future. Updated analyses similar to this one might be necessary to ascertain the current coverage of the data sources, especially if regular reports on coverage development are not issued by the sources themselves.

Other aspects related to coverage, such as the distribution of document types, language, date of publication, or indexing speed are not analysed here and could be investigated in future studies, as they are also necessary for users who need to decide which data source(s) are most suitable for their needs.

### Comparison with previous studies

The results generally agree with previous studies comparing the coverage of Microsoft Academic and Dimensions. Similarly to Harzing & Alakangas (2017a, b) and Thelwall (2017), Microsoft Academic detected more citations than WoS and Scopus. This citation detection advantage seems to be higher in the Humanities, Social Sciences, and Business & Economics than in the other areas, where in some cases Microsoft Academic had lower coverage (Physics, Chemistry). The results here cannot be directly compared to Hug and Brändle (2017), who reported that Scopus had slightly greater coverage of journal articles than Microsoft Academic, because this study does not analyse specific document types. However, assuming that most citations come from journal articles, Microsoft Academic seems to have now surpassed Scopus in raw size, at least in the three areas mentioned above.

For Dimensions, the results also agree with those reported by Harzing (2019), who found that it had a similar or better coverage than WoS and Scopus in Business & Economics. Here the results show that the three data sources offer a similar coverage (Scopus is slightly larger, followed by Dimensions), but each can detect a non-negligible proportion of citations that the others can't.

From Visser et al. (2020) the percentage of documents covered by Scopus that are also covered by Dimensions is 78%, but in this study the percentage of citations found by Scopus that are also found by Dimensions is higher (84%). The causes of the difference between these figures is unclear, but some possibilities are (a) this study uses a sample of citations while Visser et al. use the entire collection of source documents, (b) the possibility that Dimensions has a lower coverage of older documents (our study analyses citations from 2006 to 2018, while Visser et al. analysed coverage between 1996 and 2017), or (c) that there was an increase in coverage between the time Visser et al. obtained their data (December 2018), and the time the data for this study was extracted (May–June 2019). The overlap Visser et al. found between Scopus and WoS is significantly lower than found here: according to their results (overlap of 29.1 million documents, and 44.9 million documents in total in Scopus), WoS covered 65% of the documents available in Scopus. In the current study, however, WoS found 83% of the citations found by Scopus. The cause of this significant difference is also unknown, but it might be in part caused by the fact that Visser et al. analysed only documents in the SCI, SSCI, and A&HCI and the Conference Proceedings Citation Index (CPCI), while this study also considers other citation indexes within WoS Core Collection, such as ESCI and BKCI. Lastly, the percentage of Scopus documents also covered by Microsoft Academic reported by Visser et al. (81%) is very similar to the percentage of Scopus citations also found by Microsoft Academic reported here (82%). However, the full overlap between the two sources is much higher here (66%) than in Visser et al. (18%), because in the latter study a much higher amount of unique content was detected in Microsoft Academic. One possible reason for this might be that our study only considers documents with recorded relationships to other documents (through citations), while some of the documents in Microsoft Academic analysed in Visser et al. might not have these connections, which would make them undetectable to our methodology.

Although most of the results of the overlap analysis reported here closely match those of the previous study with the same seed set (Martín-Martín et al. 2018), several discrepancies were found. In two subject categories (Psychology, and Astronomy & Astrophysics), the updated analysis showed that Google Scholar had a lower coverage than the other data sources, while in the old dataset, this was not the case. In the case of Astronomy & Astrophysics, this apparent fluctuation in coverage is consistent with an editorial published in August of 2019 in the journal Astronomy & Astrophysics, which denounced a sharp decrease in the h5-index of this journal in the 2019 edition of Google Scholar Metrics (Forveille 2019) caused by a technical error in Google Scholar. Therefore, this seems to be a new case of a major coverage outage in Google Scholar, similar to one previously reported by Delgado López-Cózar and Martín-Martín (2018) which affected many journals published in Spain, and which was resolved when Google Scholar rebuilt its index a few months later. This issue will be analysed in detail in a future study as an example of how coverage in Google Scholar can suffer large (downward) fluctuations over time, as this can negatively affect literature search.

## Conclusions

### Comprehensiveness of data sources across subject categories

The results show that Google Scholar is still the most comprehensive data source among the six studied here. This holds true for the overall results and the results across all subject areas, with some exceptions such as *Astronomy & Astrophysics*. Google Scholar found nearly all the citations found by Microsoft Academic, Dimensions, and COCI (89%, 93%, and 94%, respectively). The largest divergences occur in the Humanities and Social Sciences (lowest value is 84%, which corresponds to the percentage of Scopus citations in the Humanities found by Google Scholar). Additionally, there is a significant amount of extra coverage in Google Scholar that is not found in any of the other data sources (26% of all citations across all data sources). Google Scholar could therefore make an important contribution to the scientific community by opening its bibliographic and citation data, which would also facilitate the identification of errors such as coverage fluctuations.

Whilst the results confirm that Microsoft Academic and Dimensions provide at least as many citations as Scopus and WoS in many subject categories, some gaps still exist:

- Microsoft Academic seems to index the Humanities, Social Sciences, and Business, Economics & Management particularly well, although not for all categories.
- Dimensions is closely behind Scopus in all areas in terms of citations found, but surpasses WoS in all areas, except in two (Physics & Mathematics, and Chemical & Material Sciences) where they are tied, although there are also differences at the level of subject categories (Dimensions also has coverage gaps in some Humanities categories).

### Implications for academic literature search

Although Google Scholar and Microsoft Academic are the two most comprehensive bibliographic data sources analysed in this study, their search functionalities have a number of limitations, such as limited support of Boolean and other types of search operators, limited filtering capabilities (Google Scholar), and non-transparent algorithms to process queries and rank the documents in the results page (Microsoft Academic uses artificial intelligence, and Google Scholar uses publicly unknown heuristics to rank documents by relevance) (Beel and Gipp 2009a, b, c; Martin-Martin et al. 2017; Orduña-Malea et al. 2016; Rovira et al. 2019; Wang et al. 2020). These characteristics, which prevent users from being able to generate complex search equations that are guaranteed to stay reproducible over time, have led some authors to consider Google Scholar and Microsoft Academic inadequate for query-based search (Gusenbauer and Haddaway 2020). Dimensions, which does not allow complex Boolean searches in its web interface either, was not analysed in that study.

On the other hand, Scopus and WoS have a lower coverage, especially in some areas such as the Humanities and Social Sciences, do not cover non-peer-reviewed scientific documents (Martín-Martín et al. 2018), are slower at indexing (Moed et al. 2016), and are not free. These characteristics reduce their usefulness in situations where fast and unrestricted access to the latest studies is important, such as the COVID-19 pandemic in which preprints play a critical role (Fraser et al. 2020). Nevertheless, these sources offer advanced search and filtering functionalities, and were considered suitable tools for evidence synthesis in the form of systematic reviews (Gusenbauer and Haddaway 2020).

Thus, there seems to be a mismatch between the bibliographic data sources that are currently the most comprehensive, and those that offer users the most control over their searches. Since systematic reviews require both comprehensiveness of coverage and control over the search process, it is possible that in some cases no single currently available data source is adequate for the task, and instead at least two sources should be used. One way to do this would be to expand the concept of systematic search beyond the traditional search query to include other non-query-based search processes that can also be carried out in a systematic and reproducible manner. One possibility would be the expansion of a document collection obtained in a query-based search through the analysis of its citation network. This expansion can be carried out in a more comprehensive data source, different from the one where the initial search was carried out. As a longer term solution, academic search tools should strive to offer more a transparent and reproducible search process and embrace community standards for interoperability and reuse of document metadata (Haddaway and Gusenbauer 2020).

Lastly, searches suitable for systematic reviews are only one of the many types of search that are carried out in these data sources. Indeed, the more recent academic search platforms (Microsoft Academic, Semantic Scholar, Dimensions) have not implemented traditional advanced query-based capabilities (Dimensions supports them in its API), and seem to be instead focusing on the browsing experience (advanced filtering), and in offering analytics dashboards. Lens.org seems to be an exception, as it also offers advanced structured query-based search (Tay 2019). Future studies could focus on the suitability of these and other bibliographic data sources to solve specific types of information needs, as it is important that researchers are aware of the strengths and limitations of each data source for specific use cases and in specific knowledge domains.

## Implications for bibliometric analyses

As new sources of bibliographic data (including citation data) become openly available and validated for specific types of bibliometric analyses, the need to rely on expensive proprietary data sources may diminish. Regarding the findings in this study, the final decision about which source to use may depend on properties of the sources other than coverage, such as metadata quality and bulk access options. If these factors are not of overriding importance, however, then Google Scholar is the best choice in almost all subject areas for those needing the most comprehensive citation counts but not needing complete lists of citing sources. If complete lists are needed, then Microsoft Academic is the best alternative and is also free. The amount of citation data in the public domain (through COCI) is still low and not useful on its own, presumably because its role is to feed other sources, not to be more comprehensive than them.

In use cases where exhaustiveness of coverage is required, but coverage divergence is considered to be large (many unique citations in each data source), the combination of several sources is recommended.

To conclude, the evidence presented in this study is intended to serve as a tool for researchers and other users of bibliographic databases, one that will hopefully help them make more informed decisions when they need to select one or more of these data sources to solve a specific information need.

**Acknowledgements** We thank Medialab UGR (Universidad de Granada) for providing funding to cover the cost of hosting the interactive web application[54] created to explore the data used in this study. We thank Digital Science for providing free access to the Dimensions API. We thank Jing Xuan Xie for translating the abstract to Chinese. We thank Asura Enkhbayar for suggesting the use of an upset plot in Fig. 2. Lastly, we thank two anonymous reviewers for their thoughtful comments, which have helped improved the manuscript substantially.

## Appendix 1: Complete list of Venn diagrams computed for this study

| No subject aggregation | |
|---|---|
| Two-set Venn diagrams (all subject categories) | https://osf.io/bwpaq/ |
| Three-set Venn diagrams (all subject categories) | https://osf.io/jkrge/ |
| **Aggregated by 8 subject areas** | |
| Google Scholar–Microsoft Academic–Scopus | https://osf.io/h7m8s/ |
| Google Scholar–Microsoft Academic–Dimensions | https://osf.io/7v4kr/ |
| Google Scholar–Microsoft Academic–Web of Science | https://osf.io/fn3yh/ |
| Google Scholar–Microsoft Academic–COCI | https://osf.io/s3bmp/ |
| Google Scholar–Scopus–Dimensions | https://osf.io/q8ecx/ |
| Google Scholar–Scopus–Web of Science | https://osf.io/qkc2a/ |
| Google Scholar–Scopus–COCI | https://osf.io/mrvdb/ |
| Google Scholar–Dimensions–Web of Science | https://osf.io/nwm83/ |
| Google Scholar–Dimensions–COCI | https://osf.io/dzs5x/ |
| Google Scholar–Web of Science–COCI | https://osf.io/64chg/ |
| Microsoft Academic–Scopus–Dimensions | https://osf.io/hgzn6/ |
| Microsoft Academic–Scopus–Web of Science | https://osf.io/f7xpa/ |
| Microsoft Academic–Scopus–COCI | https://osf.io/c6tpz/ |
| Microsoft Academic–Dimensions–Web of Science | https://osf.io/f5zxs/ |
| Microsoft Academic–Dimensions–COCI | https://osf.io/ry87a/ |
| Microsoft Academic–Web of Science–COCI | https://osf.io/vxyj4/ |
| Scopus–Dimensions–Web of Science | https://osf.io/xqg3y/ |
| Scopus–Dimensions–COCI | https://osf.io/jmvb6/ |
| **Aggregated by 252 subject categories (zipped)** | |
| Google Scholar–Microsoft Academic | https://osf.io/v4ek3/ |
| Google Scholar–Scopus | https://osf.io/umsyw/ |
| Google Scholar–Dimensions | https://osf.io/jqmuy/ |
| Google Scholar–Web of Science | https://osf.io/4b8uq/ |
| Google Scholar–COCI | https://osf.io/gytuh/ |
| Microsoft Academic–Scopus | https://osf.io/jw2bt/ |
| Microsoft Academic–Dimensions | https://osf.io/a2mp7/ |
| Microsoft Academic–Web of Science | https://osf.io/2hkxq/ |
| Microsoft Academic–COCI | https://osf.io/ch4gb/ |

[54] https://albertomartin.shinyapps.io/citation_overlap_2019/.

| Aggregated by 252 subject categories (zipped) | |
|---|---|
| Scopus–Dimensions | https://osf.io/q4swk/ |
| Scopus–Web of Science | https://osf.io/qcpbh/ |
| Scopus–COCI | https://osf.io/2xvbh/ |
| Dimensions–Web of Science | https://osf.io/pdycb/ |
| Dimensions–COCI | https://osf.io/j7qte/ |
| Web of Science–COCI | https://osf.io/mnwe7/ |

## Appendix 2: Complete list of boxplots computed for this study

| Subject category-level overlap data aggregated by 8 subject areas | |
|---|---|
| Google Scholar–Microsoft Academic | https://osf.io/b94xp/ |
| Google Scholar–Scopus | https://osf.io/rvbw9/ |
| Google Scholar–Dimensions | https://osf.io/ubtrm/ |
| Google Scholar–Web of Science | https://osf.io/7wb49/ |
| Google Scholar–COCI | https://osf.io/7ekdr/ |
| Microsoft Academic–Scopus | https://osf.io/jx7by/ |
| Microsoft Academic–Dimensions | https://osf.io/x4257/ |
| Microsoft Academic–Web of Science | https://osf.io/rdw7g/ |
| Microsoft Academic–COCI | https://osf.io/f8a9e/ |
| Scopus–Dimensions | https://osf.io/3a97k/ |
| Scopus–Web of Science | https://osf.io/w4zv3/ |
| Scopus–COCI | https://osf.io/jtnyu/ |
| Dimensions–Web of Science | https://osf.io/gsjwm/ |
| Dimensions–COCI | https://osf.io/sr4wu/ |
| Web of Science - COCI | https://osf.io/6dkw4/ |

## References

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies, 1*(1), 377–386. https://doi.org/10.1162/qss_a_00019.

Beel, J., & Gipp, B. (2009a). Google Scholar's ranking algorithm: The impact of articles' age (an empirical study). *Sixth International Conference on Information Technology: New Generations, 2009,* 160–164. https://doi.org/10.1109/ITNG.2009.317.

Beel, J., & Gipp, B. (2009b). Google Scholar's ranking algorithm: The impact of citation counts (An empirical study). *Third International Conference on Research Challenges in Information Science, 2009,* 439–446. https://doi.org/10.1109/RCIS.2009.5089308.

Beel, J., & Gipp, B. (2009c). Google Scholar's ranking algorithm: An introductory overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)* (pp. 230–241). http://www.issi-society.org/proceedings/issi_2009/ISSI2009-proc-vol1_Aug2009_batch2-paper-1.pdf

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies, 1*(1), 363–376. https://doi.org/10.1162/qss_a_00018.

Chapman, K., & Ellinger, A. E. (2019). An evaluation of Web of Science, Scopus and Google Scholar citations in operations management. *The International Journal of Logistics Management, 30*(4), 1039–1053. https://doi.org/10.1108/IJLM-04-2019-0110.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171–176. https://doi.org/10.1145/363958.363994.

Delgado López-Cózar, E., & Martín-Martín, A. (2018). Apagón digital de la producción científica española en Google Scholar. *Anuario ThinkEPI, 12,* 265–276. https://doi.org/10.3145/thinkepi.2018.40.

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators*. Berlin: Springer.

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., & Parsonage, H. (2018). *data.table: Extension of 'data.frame'* (1.11.4).

Else, H. (2018, April 11). How I scraped data from Google Scholar. *Nature*. https://doi.org/10.1038/d41586-018-04190-5

Forveille, T. (2019). A&A ranking by Google. *Astronomy & Astrophysics, 628,* E1. https://doi.org/10.1051/0004-6361/201936429.

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., & Coates, J. A. (2020). Preprinting a pandemic: The role of preprints in the COVID-19 pandemic. *BioRxiv*, 2020.05.22.111294. https://doi.org/10.1101/2020.05.22.111294

Gusenbauer, M. (2018). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*. https://doi.org/10.1007/s11192-018-2958-5.

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217. https://doi.org/10.1002/jrsm.1378.

Haddaway, N., & Gusenbauer, M. (2020, February 3). A broken system: Why literature searching needs a FAIR revolution. *Impact of Social Sciences*. https://blogs.lse.ac.uk/impactofsocialsciences/2020/02/03/a-broken-system-why-literature-searching-needs-a-fair-revolution/.

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). 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, 11*(3), 823–834. https://doi.org/10.1016/J.JOI.2017.06.005.

Harzing, A. W. (2016). Microsoft Academic (Search): A Phoenix arisen from the ashes? In *Scientometrics* (Vol. 108, No. 3, pp. 1637–1647). Springer, Netherlands. https://doi.org/10.1007/s11192-016-2026-y

Harzing, A.-W. (2016). Sacrifice a little accuracy for a lot more comprehensive coverage. *Harzing.Com*. https://harzing.com/blog/2016/08/sacrifice-a-little-accuracy-for-a-lot-more-comprehensive-coverage

Harzing, A. W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? In *Scientometrics* (Vol. 120, Issue 1, pp. 341–349). Springer, Netherlands. https://doi.org/10.1007/s11192-019-03114-y

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9.

Harzing, A. W., & Alakangas, S. (2017a). Microsoft Academic: Is the phoenix getting wings? In *Scientometrics* (Vol. 110, Issue 1, pp. 371–383). Springer, Netherlands. https://doi.org/10.1007/s11192-016-2185-x

Harzing, A. W., & Alakangas, S. (2017b). Microsoft Academic is one year old: The Phoenix is ready to leave the nest. In *Scientometrics* (Vol. 112, Issue 3, pp. 1887–1894). Springer, Netherlands. https://doi.org/10.1007/s11192-017-2454-3

Haunschild, R., Hug, S. E., Brändle, M. P., & Bornmann, L. (2018). The number of linked references of publications in Microsoft Academic in comparison with the Web of Science. In *Scientometrics* (Vol. 114, Issue 1, pp. 367–370). Springer, Netherlands. https://doi.org/10.1007/s11192-017-2567-8

Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*. https://doi.org/10.1007/s11192-019-03217-6.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies, 1*(1), 414–427. https://doi.org/10.1162/qss_a_00022.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies, 1*(1), 387–395. https://doi.org/10.1162/qss_a_00020.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics, 3,* 23. https://doi.org/10.3389/frma.2018.00023.

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., et al. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00031.

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics, 113*(3), 1551–1571. https://doi.org/10.1007/s11192-017-2535-3.

Kousha, K., & Thelwall, M. (2018). Can Microsoft Academic help to assess the citation impact of academic books? *Journal of Informetrics, 12*(3), 972–984. https://doi.org/10.1016/j.joi.2018.08.003.

Kousha, K., Thelwall, M., & Abdoli, M. (2018). Can Microsoft Academic assess the early citation impact of in-press articles? A multi-discipline exploratory analysis. *Journal of Informetrics, 12*(1), 287–298. https://doi.org/10.1016/j.joi.2018.01.009.

Krassowski, M. (2020). *ComplexUpset*. https://github.com/krassowski/complex-upset

Larsson, J., Godfrey, A. J. R., Kelley, T., Eberly, D. H., Gustafsson, P., & Huber, E. (2018). *eulerr: Area-Proportional Euler and Venn Diagrams with Circles or Ellipses* (4.1.0).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Martín-Martín, A. (2018). *Code to extract bibliographic data from Google Scholar* (v1.0). Zenodo. https://doi.org/10.5281/zenodo.1481076

Martín-Martín, A., & Delgado López-Cózar, E. (2016). *Reading Web of Science data into R* (0.6).

Martin-Martin, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics, 11*(1), 152–163. https://doi.org/10.1016/j.joi.2016.11.008.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics, 12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002.

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics, 10*(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017.

Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: Re-discovering the ecosystem of scientific information. *Profesional de La Informacion, 27*(2), 420–431. https://doi.org/10.3145/epi.2018.mar.21.

Orduña-Malea, E., Martín-Martín, A., Ayllon, M., & Delgado López-Cózar, E. (2014). The silent fading of an academic search engine: The case of Microsoft Academic Search. *Online Information Review, 38*(7), 936–953. https://doi.org/10.1108/OIR-07-2014-0169.

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar: Destapando la caja de Pandora académica*. Universidad de Granada y Unión de Editoriales Universitarias Españolas.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. *Revista Española de Documentación Científica, 40*(4), e185. https://doi.org/10.3989/redc.2017.4.1500.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). Classic papers: Using Google Scholar to detect the highly-cited documents. In *23rd International conference on science and technology indicators* (pp. 1298–1307). https://doi.org/10.31235/osf.io/zkh7p

Ortega, J. L. (2014). *Academic search engines: A quantitative outlook*. Cambridge: Chandos Publishing.

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies, 1*(1), 428–444. https://doi.org/10.1162/qss_a_00023.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*.

Rovira, C., Codina, L., Guerrero-Solé, F., & Lopezosa, C. (2019). Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft academic, WoS and scopus. *Future Internet, 11*(9), 202. https://doi.org/10.3390/fi11090202.

Shotton, D. (2013). Publishing: Open citations. *Nature, 502*(7471), 295–297. https://doi.org/10.1038/502295a.

Shotton, D. (2018). Funders should mandate open citations. *Nature, 553*(7687), 129. https://doi.org/10.1038/d41586-018-00104-7.

Tay, A. (2019, April 3). 6 reasons why you should try Lens.org. *Medium*. https://medium.com/@aarontay/6-reasons-why-you-should-try-lens-org-c40abb09ec6f

Thelwall, M. (2017). Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. *Journal of Informetrics, 11*(4), 1201–1212. https://doi.org/10.1016/j.joi.2017.10.006.

Thelwall, M. (2018a). Does Microsoft Academic find early citations? *Scientometrics, 114*(1), 325–334. https://doi.org/10.1007/s11192-017-2558-9.

Thelwall, M. (2018b). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics, 12*(1), 1–9. https://doi.org/10.1016/j.joi.2017.11.001.

Thelwall, M. (2018c). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics, 12*(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006.

van der Loo, M., van der Laan, J., R Core Team, Logan, N., & Muir, C. (2018). *stringdist: Approximate String Matching and String Distance Functions* (0.9.5.1).

van Eck, N. J., & Waltman, L. (2019). *Accuracy of citation data in Web of Science and Scopus*.

van Eck, N. J., Waltman, L., Larivière, V., & Sugimoto, C. (2018). *Crossref as a new source of citation data: A comparison with Web of Science and Scopus*. https://www.cwts.nl/blog?article=n-r2s234&title=crossref-as-a-new-source-of-citation-data-a-comparison-with-web-of-science-and-scopus

Van Noorden, R. (2014). November 7). *Google Scholar pioneer on search engine's future. Nature.*. https://doi.org/10.1038/nature.2014.16269.

Visser, M., van Eck, N. J., & Waltman, L. (2020). *Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic*. https://arxiv.org/abs/2005.10732

Walker, A., & Braglia, L. (2018). *openxlsx: Read, Write and Edit XLSX Files* (4.1.0).

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies, 1*(1), 396–413. https://doi.org/10.1162/qss_a_00021.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*.

Wu, J., Kim, K., & Giles, C. L. (2019). CiteSeerX: 20 years of service to scholarly big data. *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*. https://doi.org/10.1145/3359115.3359119.

International Business Review xxx (xxxx) xxxx

Contents lists available at ScienceDirect

International Business Review

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

# The art of writing literature review: What do we know and what do we need to know?

Justin Paul[a,b,*], Alex Rialp Criado[c,d]

[a] University of Puerto Rico, San Juan, PR, USA
[b] Indian Institute of Management (IIM-K), India
[c] Universitat Autònoma Barcelona, Spain
[d] Norwegian University of Science and Technology, Norway

ABSTRACT

A literature review article provides a comprehensive overview of literature related to a theme/theory/method and synthesizes prior studies to strengthen the foundation of knowledge. In the growing International Business (IB) research field, systematic literature reviews have great value, yet there are not many reviews published describing how researchers can design and develop classic review articles. In explaining the purpose, methodology, and structure of a systematic review, we provide guidelines for developing most insightful and useful review articles. By outlining steps and thumb rules to keep in mind, we present an overview of different types of review articles and explain how future researchers could potentially find them useful. In addition, we introduce nine articles finally selected for this special issue of systematic literature review-Looking back to look forward International Business research in the days to come.

## 1. Introduction

A subject advances when prior studies are synthesized logically based on the findings of prior studies (Kumar, Paul, & Unnithan, 2019). Literature reviews, as a research methodology (Snyder, 2019), contribute significantly for conceptual, methodological, and thematic development of different domains (Palmatier, Houston, & Hulland, 2018; Hulland & Houston, 2020). *Review papers* "are critical evaluations of prior studies that have already been published" (Bem, 1995, p. 172). They include, among others, systematic reviews and meta-analytical reviews exploring quantitative effects. Review articles carefully identify and synthesize relevant literature to compare and contrast the findings of prior studies in a domain. Thus, review articles provide readers with a state-of-the-art understanding of the research topic, help identify research gaps and signal future research avenues. In other words, systematic reviews, in particular, provide a number of critical discussions on a specific research theme by integrating extant literature, synthesizing prior studies, identifying knowledge gaps, and developing new theoretical frameworks (Marabelli & Newell, 2014). Systematic reviews, in particular, have become an explicitly recognized form of review-based research in many different disciplines (Callahan, 2014, p. 272; Kraus, Breier, & Dasí-Rodríguez, 2020). Many journals such as Journal of Management, Journal of International Business Studies, Journal of Organizational Behavior, Journal of Academy of Marketing Science, International Journal of Consumer Studies etc. have launched annual special issues for review articles in the recent years. In addition, there are exclusive journals publishing review articles such as International Journal of Management Reviews (IJMR), Journal of Economic Literature (JEL) and Academy of Management Review.

It is worth noting that hundreds of research papers have been published using the same old theories, measures, and methods. One of the important goals of a review article is to identify key research gaps based on what constructs, theories and methods are widely applied in different settings and in what contexts (industry as well as country) studies have been carried out. Accordingly, authors of a classic review article provide directions for future research with reference to new and novel ideas, theories, measures, methods and novel research questions. Thus, a review article can serve as a platform for future research. They set the goal to discourage researchers from using the same old theories and methods in a recycled and replete way. A very well crafted literature review article has the potential to serve as base/platform/lens/springboard for future research because such an article explicitly synthesizes current knowledge, identifies research gaps, and suggests exciting new directions for future research in a given field of research, with reference to Methodology, Constructs/Variables, Theory and Contexts. Similarly, theoretical models developed as part of literature review studies can be used by both researchers and practitioners as typologies/base/lens in their research studies using quantitative or qualitative methods and/or practice. Therefore, once published, they would/usually become a very welcome and great addition to the literature.

* Corresponding author.
*E-mail addresses:* profjust@gmail.com (J. Paul), Alex.Rialp@uab.cat (A.R. Criado).

https://doi.org/10.1016/j.ibusrev.2020.101717

0969-5931/ © 2020 Elsevier Ltd. All rights reserved.

Please cite this article as: Justin Paul and Alex Rialp Criado, International Business Review, https://doi.org/10.1016/j.ibusrev.2020.101717

ARTICLE IN PRESS



## 2. Methodology and structure of review articles

Systematic literature review articles can be broadly classified as domain-based, theory-based, and method-based. In addition to these categories of systematic literature reviews, meta analytical reviews are also increasingly popular in many different subject areas (Hulland & Houston, 2020). There are large number of domain-based reviews available in almost all subject areas both business-related (management, marketing, finance and accounting, entrepreneurship, etc.) and non-business related. However, there are not many well-crafted theory-based and method-based reviews published in well recognized journals.

### 2.1. Domain-based review

Domain-based review articles can be classified into different categories. Namely – Structured review focusing on widely used methods, theories and constructs (Canabal & White, 2008; Kahiya, 2018; Paul & Singh, 2017; Paul & Feliciano-Cestero, 2020; Rosado-Serrano, Paul, & Dikova, 2018), Framework-based (Paul & Benito, 2018), Bibliometric review (Randhawa, Wilden, & Hohberger, 2016), Hybrid-Narrative with a framework for setting future research agenda (Dabić et al., 2020; Kumar et al., 2019; Paul, Parthasarathy, & Gupta, 2017), and Review aiming for model/framework development (Paul & Mas, 2019; Paul, 2019). These classifications can be elaborated as follows.

#### 2.1.1. Structured review

When a domain-based review article is structured scientifically and specifically based on widely used methods, theories, constructs in the form of tables and figures, readers get insightful information from the data reported and content. Such information is normally presented in well designed tables in classic structured review articles. This helps especially junior researchers to understand what kind of methods have been already used and what theories and constructs have already been applied. Researchers can identify research gaps with reference to methods, theories and constructs based on the compiled information. Some of the classic review articles found in the extant literature falls in this category (Canabal & White, 2008, Kahiya, 2018; Paul & Feliciano-Cestero, 2020). This type of domain review articles usually have between 5–10 useful tables in structured format.

#### 2.1.2. Framework-based review

A domain-based review article can be called as Framework-based review if the authors develop it using a framework such as ADO (Antecedents, Decisions and Outcome), as seen, for instance, in Paul & Benito's (2018) review article, or the 6 W Framework developed by Callahan (2014). This 6 W Framework is comprised of – Who, When, Where, How, What, and Why. Xie, Reddy, and Liang (2017) demonstrated how to use this 6 W framework in a literature review article on cross-border acquisitions. Another useful framework is Theory, Construct, Characteristics and Methodology (TCCM) developed and applied by Paul and Rosado-Serrano (2019), or the 7-P Framework (Paul & Mas, 2019). Thematic reviews with a framework have proven to be more acceptable as they are likely to show a more robust structure. Therefore, authors of framework-based reviews have to either develop their own framework and use it for structuring their review or, adopt an already existing framework like ADO duly acknowledging whom they are borrowing it from if developed by others.

#### 2.1.3. Bibliometric review

Bibliometric reviews analyse an extensive amount of published research by using statistical tools, thus to figure out trends and citations and/or co-citations of a particular theme, by year, country, author, journal, method, theory, and research problem. A graphical bibliometric review can be developed using Viewer software programs currently available such as VoS (Visualization of Similarities), which is widely used to carry out such a type of bibliometric review in diverse subject areas, including International Business (Rialp, Merigó, Cancino, & Urbano, 2019). An issue inherent in many bibliometric analyses is that out of a given pool of articles, a relatively small number of articles represent a major part of the total citations in the analysis. Some researchers, however, remain somewhat sceptical regarding the overall impact of bibliometric analyses, compared to other types of reviews (Fetscherin & Heinrich, 2015). In our view, bibliometric reviews do not deal with theories, methods, and constructs as much as they usually do with authors, affiliations, countries, citations and co-citations, etc.

#### 2.1.4. Hybrid review

Hybrid reviews can be developed in, at least, two different ways: i) When researchers integrate a framework to provide directions for future research in a more narrative-oriented type of literature review, it can be called as a hybrid type review. For example, Paul, Parthasarathy and Gupta (2017) used Theory, Context and Methods (TCM) framework in their narrative type review on exporting challenges for small firms to provide directions for research. ii) A second way of a rather hybrid form of review may be developed by integrating the tenets of both bibliometric and structured reviews. Further in this special issue, Bahoo, Alon and Paltrinieri, for instance, have followed a similar approach in their review focused on corruption in international business. They integrated the tenets of bibliometric review with that of a structured review.

#### 2.1.5. Review aiming for theory development

A very significant number of review articles published in highly exclusive business journals, such as Academy of Management Review or Academy of Marketing Science Review, fall in this category. In this case, authors typically develop theoretical models and/or testable hypotheses or propositions in such theory-building review articles. However, they do not necessarily proceed to test those models and/or theoretical propositions in the same article. Paul and Mas' (2019) article on 'Toward a 7 P framework for international marketing' is a clear example for this type of work. Very recently, Post, Sarala, Gatrell, and Prescott (2020) provide a great contribution with plenty of indications and guidelines about how to advance theory by means of review articles.

### 2.2. Theory-based review

Systematic reviews focused analysing the role of a specific theory in a subject area/ field are very useful for both senior and junior researchers. Such a review article can be labelled as Theory-based review. This type of review articles synthetize and help advancing a body of literature that uses and/or empirically applies a given underlying theory. For example, Rindfleisch and Heide's (1997) classic review titled Transaction Cost analysis in Marketing: Past, Present and Future Applications' has been cited more than 2600 times. Other examples of theory-based reviews in the Marketing field are 'Resource-Based Theory in Marketing' (Kozlenkova, Samaha, & Palmatier, 2014) in the Journal of Academy of Marketing Science or the one titled 'Role of self-determination theory in marketing science' (Gilal et al., 2019). Also, a very recent review on studies employing Gradual internationalization versus Born-global models (Paul & Rosado-Serrano, 2019) falls in this category in the area of International Business/Marketing. Similar reviews can be developed exploring the role and application of a given core theory -or even different theories- in a given field (Eisenhardt, 1989), sometimes with a special emphasis on theoretical contributions and/or empirical developments in specific set of scientific journals (Colquitt & Zapata-Phelan, 2007). Further developments of this nature can also imply, for instance, systematically reviewing Agency Theory in Franchising, or the Theory of Planned Behaviour in International Business/Marketing or Entrepreneurship, etc.

ARTICLE IN PRESS



### 2.3. Method-based review

Method-based review articles synthetize and extend a body of literature that uses an underlying methodology (either quantitative or qualitative). For example, the paper titled 'Event Study Methodology in the Marketing Literature: An Overview' (Sorescu, Warren, & Ertekin, 2017). Similarly, the article 'Discriminant Validity Testing in Marketing: An Analysis, Causes for Concern, and Proposed Remedies' by Voorhees, Brady, Calantone, and Ramirez (2016) systematically reviews existing approaches for assessing discriminant validity in marketing contexts applying Monte Carlo simulation to determine which tests are most effective. However, the number of method-based reviews available in different subject areas of business administration or International Business/Entrepreneurship are not so many (some notable exceptions in the International Entrepreneurship field being, for instance, Coviello and Jones (2004) or, more recently, Ji, Plakoyiannaki, Dimitratos, and Chen (2019)). Therefore, there are still great opportunities for developing such method-based review articles. For example, review articles focusing on Smart PLS applications in global strategy research or Structural Equation Modelling (SEM) in a specialized area of International Business/Marketing empirical literature can be developed and published.

### 2.4. Meta analytical review

While both focusing mainly on examining quantity or volume of previous research, systematic reviews and meta-analysis actually differ; the former seeks to synthesize many previous findings, while the latter makes a deeper statistical assessment of available data and findings (essentially correlations among variables) from many previous quantitative studies (Pati & Lorusso, 2018; Piper, 2013). A meta-analysis is a form of increasingly popular quantitative technique that is being widely recognized as perhaps one of the best statistical assessment of prior empirical research on a specific research topic. Meta-analyses help researchers to 'identify directions and effect sizes based on prior studies with the help of weighted average techniques, and contextualize the relationships by considering moderator variables' (Klier, Schwens, Zapkau, & Dikova, 2017, p. 3??). We could also refer here to the classic meta-analytical review developed with a proper methodology and structure by Knoll and Matthes (2017), published in the Journal of Academy of Marketing Science. Similalry, Rauch, Wiklund, Lumpkin, and Frese (2009) focused on the classical relationship between entrepreneurial orientation and business performance published in Entrepreneurship Theory and Practice. Also, two articles in this Special Issue volume are meta-analytical reviews, i) by Tang and Buckley (2020) ii) by Schmid and Morschett (2020). They developed meta-analysis reviews on hard core topics in the field of International Business.

## 3. Thumb rules and suggestions for developing an impactful review article

Based on our own knowledge and experience as editors, guest editors and authors of several review articles, and partly complementing other similar efforts (Fisch & Block, 2018; Reuber, 2010; Webster & Watson, 2002), we succinctly provide some potentially useful tips and suggestions for developing more insightful and impactful review articles in future research.

### 3.1. Topic selection

Not surprisingly, well-crafted review articles tend to be generally impactful. However, authors should not select a very recurrent topic for review when there are already other excellent reviews on the same topic (especially very recent ones) published in highly reputed journals. Editors and reviewers may not be keen to consider very traditional thematic reviews when there are several comprehensive ones already available elsewhere related to a given theme/topic unless authors demonstrating a very novel reviewing contribution by providing a completely new set of research agenda. It is important then to check this thematic novelty on key bibliometric databases such as Google Scholar, Web of Science (WoS) or Scopus before deciding to choose a more generic versus specific topic for review.

### 3.2. Journal selection criteria, identification of streams and period coverage

Normally, many researchers and academics tend to select perhaps the most well-known bibliographic database, Web of Science (WoS)/Social Science Citation Index (SSCI)/Journal Citation Report which list academic journals with an Impact Factor (IF), for identifying potential sources for reviewing. When there are several hundreds of papers on a highly popular topic already published to be potentially reviewed, one can even rely upon JCR-indexed journals with an IF above a given threshold (i.e. 1.0 plus following, for instance, Paul & Rosado-Serrano, 2019). Also, many authors have published review articles using studies from the indexed journals found in Scopus, which lists a greater number of journals than WoS. Therefore, relying mainly on Scopus to conduct a systematic literature review may yield a very long list of references which may even exceed the word limits set by many journals. On the other hand, we have come across some published review articles of a relatively small size samples of articles in a specific field justifying their selection on 5–10 journals with a minimum rank of 3 star and/or above in the Journal Quality List (JQL) of the Association of Business Schools (ABS) or Journals with an A or A star rank in the Australian Business Deans Council (ABDC) list. At the same time, it is important to keep in mind that journals might not be extremely interested in your review, if it does not cover also articles from your target journal. Therefore, it is advisable to include articles from at least 10–20 significant journals in a review paper, to minimise the risk of not publishing your work due to biased journal selection criteria. Most of the review articles in this Special Issue cover articles retrieved from well-established bibliographic databases such as WoS and/or Scopus. Nonetheless, it was surprising to note that some of the mainly rejected submissions for this special issue did not have clear journal selection criteria and most of them included references from not fully reliable academic sources.

#### 3.2.1. Articles search and inclusion criteria using keywords

A systematic review article can be developed using 40–50 to 500 or more relevant papers. Sourcing relevant articles can be, however, a challenge. Authors will have to use their knowledge, judgment and experience many times for deciding upon clear selection criteria (i.e. exclusion/inclusion) of articles in their sample. There are two popular methods for determining, among others, highly convenient inclusion criteria: i) Keywords decided by the authors of a potential article to be selected for being reviewed are generally found directly in the title, abstract or list of keywords. ii) Keywords can be also found in the full text of the article, apart from in its title or abstract. Therefore, the sample size of a review article will tend to be relatively small if only the first criteria is strictly used. However, authors should be aware that they might get hundreds of papers to be potentially included in their sample, if they use second criteria including also keywords in the full text. In that case, wide reading of content, discussion, deliberation, and consensus among the author/s of a review paper is needed many times in order to decide the most appropriate final sample.

#### 3.2.2. Identification of streams and time period of the review

Several review articles focus on identifying the main sub-streams of research conducted in the past on a wider topic or even an entire discipline like, for instance, Strategic Management (Furrer, Thomas, & Goussevskaia, 2008; Hoskisson, Wan, Yiu, & Hitt, 1999). Jones, Coviello, and Tang (2011) proceeded this way in their assessment of the International Entrepreneurship field published in the Journal of

ARTICLE IN PRESS



Business Venturing. More recently, Dabić et al. (2020) identify different streams of past research on immigrant entrepreneurship. Another review on Social entrepreneurship published in Journal of Business Research (Gupta, Chauhan, Paul, & Jaiswal, 2020), and a review on Culture and International Business by Srivastava, Singh, and Dhir (2020), included in this Special Issue of International Business Review (IBR), also identify and provide a clear overview of the sub-streams of research in their specific fields.

As regards the time coverage of a review paper, it can be found that some reviews cover just (or less than) 10 years while there are other reviews covering up to 50 years or more of prior research in the field (Paul & Feliciano-Cestero, 2020; Schmid & Kotulla, 2011). Review articles covering 20, 25 or 30 years of research are also relatively common (Furrer et al., 2008). In our opinion, it is important to cover at least a bare minimum of a 10 year period for a systematic literature review (Rialp, Rialp, & Knight, 2005).

### 3.3. Appropriate title

Writing an integrative literature review actually implies using past and present research to explore the future (Torraco, 2016; Webster & Watson, 2002). Therefore, it is paramount to mention that beyond covering past and current research lines, the main goal of an outstanding review article is also to provide detailed and specific directions for future research. Therefore, ideally, this objective should be quite explicit and/or included in the paper's title (Rialp, Rialp, & Knight, 2014; Rauch et al., 2009; Rialp et al., 2005). For example, some of the most popular review articles of the guest editors of the Special Issue are titled as follows: 'Masstige Marketing: A review, synthesis and research agenda' (Kumar et al., 2019); 'International Franchising: A review and future research agenda' (Rosado-Serrano et al., 2018); 'Marketing in Emerging countries, A review, theoretical synthesis and extension' (Paul, 2019). A relative short title highlighting both a thorough review effort (looking back) and developing a future research agenda (looking forward) is more attractive if the researcher is aimed at focusing not only on reviewing prior research in the field but also on providing meaningful directions for future research with reference to (new) theory, methods, and constructs.

### 3.4. research gaps and importance of directions for future research

Authors are required to identify key research gaps in a good review article based upon a thorough coverage of prior research. Therefore, At least 20–25 % of the review paper, should be dedicated to develop a comprehensive future research agenda with reference to theory, methodology, constructs, and/or context. Authors need to list out and anticipate the underexplored theories, key constructs and potentially novel methods that can be used in future research in this particular but highly relevant section of a review article. Significantly, all the review studies finally selected for inclusion and publication in this Special Issue of International Business Review (IBR) carefully included a dedicated section on directions for future research.

### 3.5. Tables/figures

Authors need to understand well how tables and figures tend to be crafted, designed and/or structured in classic, most downloaded review articles. They should think twice whether such tables, charts or figures to be potentially inserted in a review article are indeed useful for others or not, essentially by thinking carefully about how many are needed to use and how to better design them. It is very recommended to look carefully at these graphical resources as included in other outstanding review articles in the field and decide if you want to add/delete some tables or figures to help the reader to better interpret them. For example, three particular recent reviews (Hao et al., 2019; Kahiya, 2018; Paul & Feliciano-Cestero, 2020) offer several well-structured tables with useful data and synthesizing content for readers and other researchers.

### 3.6. And above all: rigor and relevance

Review manuscripts are supposed to thoroughly synthesize a significant and important research area. Many times, authors have good and relevant topics. However, they fail to demonstrate well what general or more specific theories, constructs and methods are widely used and most researched. Unfortunately, many authors do not take enough efforts to pool the findings of prior studies in the best possible way. Ideally, pooled findings of prior studies need to be also reported in a table/chart format, categorising similar or contradictory findings. Also, authors of review papers have to rigorously complement text and tables regarding the most widely used methods, theories, variables, and extensively studied industry contexts, countries, etc. Undoubtedly, reviews structured both scientifically and logically, and especially showing very useful outcomes for readers are likely to be more rigorous, relevant and impactful.

## 4. Looking back to look forward: generalizations in international business research

In this section, we introduce the nine papers selected for this Special Review Issue based on competitive review process out of 76 submissions received in response to our special issue call for papers. Due to the bulk of submissions to be managed, the two guest editors assumed approximately half of the submissions each one, and took full responsibility for their management throughout the review process separately (including reviewers' selection and multiple interactions with both contributors and reviewers), with a final editorial coordination and joint agreement regarding those finally selected for this Special Review Issue of International Business Review (IBR). All papers were reviewed by three or four reviewers.

### 4.1. Cognitive foundations of firm internationalization: a systematic review and agenda for future research (Niittymies & Pajunen, 2020)

Niittymies and Pajunen address the fundamental role of managerial cognition in the internationalization of firms. However, according to these authors, there exists no coherent understanding of how prior research has examined and captured the cognitive foundations of internationalization. Niittymies and Pajunen's review identifies three mainstreams of research that, overall, consists of nine more specific research areas. They also show that especially the areas addressing (1) managerial learning, (2) characteristics of upper echelons, (3) intra-organizational perceptions, and (4) external actors' perceptions provide opportunities for the further advancement of internationalization literature. For harnessing these opportunities, those authors believe that the micro foundational approach could support the empirical examination of the cognitive foundations and would notably contribute to the Uppsala model-based theorization of the firm internationalization process.

### 4.2. Piecing together a puzzle—a review and research agenda on internationalization and the promise of exaptation (Aaltonen, 2020)

Aaltonen's review illustrates the commonalities between research agendas in the internationalization process and provides a starting point for subsequent theory development utilizing exaptation in predicting internationalization. Thus, her review contributes to the field of International Business by offering a conceptual framework to combine internationalization theories by including non-linear, discontinuous, and novel events more tightly to the existing foundations of internationalization. This is a framework-based review using TCCM protocol developed by Paul and Rosado-Serrano (2019). According to Aaltonen,

exaptation (seen as discontinuous developmental shifts) and adaptive behaviour are both Darwinian concepts used in organizational behaviour theories. Organizational behaviour also forms the basis of several internationalization theories, and exaptation is suggested to provide a theoretical tool for understanding disruptive development in internationalization. Together with adaptation, the concept illustrates a joint framework for understanding both disruptive and non-disruptive development in internationalization.

### 4.3. Corruption in international business: a review and research agenda (Bahoo, Alon, & Paltrinieri, 2020)

In their hybrid type review, combining the elements of bibliometric and structured review, Bahoo et al. (2020) systematically review the literature on the topic of corruption in International Business (137 articles) for the last 17 years between 1992 and 2019. Additionally, they identify seven research streams in this growing literature stream: (1) the legislation against corruption, (2) the determinants of corruption, (3) combating corruption, 4) the effect of corruption on firms, (5) the political environment and corruption, (6) corruption as a challenge to existing theories of management, and (7) the effect of corruption on foreign direct investment and trade. Based on their systematic review, these authors recommend that strong international laws are needed to minimize the negative impact of corruption on International Business. Firms must also consider corruption when formulating strategies to increase operational efficiency and performance. Finally, corruption challenges some key assumptions of existing theories of management. They have developed several research questions for future research in the area of International Business.

### 4.4. Export market orientation: an integrative review and directions for future research. (İpek & Bıçakcıoğlu-Peynirci, 2020)

A firm's export market orientation has long been the interest of several scholars and has received theoretical and empirical research attention in the International Business literature. In this context, İpek and Bıçakcıoğlu-Peynirci (2019)'s contribution critically investigates and synthesizes the empirical body of research on the export market orientation phenomenon in relation to theoretical issues, context, conceptual approaches, and interrelationships among the constructs of interest, and methodology. Within the scope of this systematic review, 80 studies on export market orientation published between 1998 and 2018 are subjected to a content-analysis. The findings delineate that despite the significant progress achieved in the knowledge of export market orientation, particular concerns should be still addressed to make the export market orientation literature move toward maturity.

### 4.5. Decades of research on foreign subsidiary divestment: what do we really know about its antecedents? (Schmid & Morschett, 2020)

Research on the antecedents of foreign subsidiary divestment has grown in the last several decades. However, the findings are ambiguous. Schmid and Morschett try to clarify this situation by providing, for 18 antecedent candidates derived from 45 articles, a descriptive picture of previous studies, theoretical arguments for the expected direction of effect, and quantitative synthesis of the effects by means of meta-analysis. According to this meta-analytic contribution, ten variables significantly affect the likelihood of foreign divestment while the effects of eight antecedents are inconclusive. Overall, subsidiary level antecedents have stronger effects on the divestment likelihood than parent firm or host country characteristics. According to Schmid and Morschett's findings, the resource-based view and the transaction cost approach appear to provide better explanations for foreign divestment than organizational learning theory or institutional theory. For the future research agenda, the authors propose investigating strategic motivations, taking a portfolio perspective, testing full conceptual models, considering multilevel data structures, and using Boddewyn's reversed eclectic paradigm as theoretical framework.

### 4.6. Host country risk and foreign ownership strategy: meta-analysis (Tang & Buckley, 2020)

Empirical evidence for the relationship between host country risk and a firm's ownership level in its foreign entry strategy is, according to Tang and Buckley, inconclusive. These authors revisit this relationship by integrating the internalization logic with an institution-based view to examine the moderating effects of formal and informal institutions in the home country. By meta-analysing 64 empirical studies involving 52,229 ownership decisions on foreign market entry, their study gives support to theoretical arguments that the focal relationship is positively moderated by institutional constraints on policymakers and risk-taking tendencies in the home country, but is negatively moderated by the joint effect of these two institutional factors. Tang and Buckley's meta-analytic findings shed new light on the literature of host country risk and foreign ownership strategy. Besides describing the implications of the findings for theory and practice, they also discuss the agenda for future theory development in the International Business field.

### 4.7. Foreign location decisions through an institutional lens: a systematic review and future research agenda (Donnelly & Manolova, 2020)

In their article, Donnelly and Manolova (2020) address one of the most relevant strategic decision in International Business (IB),- selection or choice of foreign location.. While there is general agreement that institutions influence location decisions, less is known, according to them, about the specific levels and mechanisms of institutional influence. To address these gaps, these authors systematically review and synthesize 106 articles published in 19 general management and IB journals from 1998 to 2019. They examine institutions at different levels (e.g. regional, national, or subnational). The characteristics and experiences of multinational corporations are deeply examined, as well as the industry conditions that determine the boundaries of institutional influence. Key findings from Donnelly and Manolova's descriptive and thematic analyses reveal both theoretical tensions and empirical gaps. Using an organizing framework, they outline four main research avenues are also identified.

### 4.8. The determinants and performance of early internationalizing firms: a literature review and research agenda (Jiang, Kotabe, Zhang, Hao, & Wang, 2020)

As scholars have examined the antecedents, processes, and performance of early internationalizing firms in the past three decades, the domain has become a full-fledged research field. However, extant reviews have not yet provided a comprehensive picture of the determinants of early internationalizing firms and their performance although it is a relevant topic in the literature. In response, Jiang et al. (2020)'s article seeks to systematically review and synthesize extant research on the determinants and performance of early internationalizing firms. The authors critically assess and examine 167 articles that have appeared in 28 academic journals over the last three decades. This study contributes to the extant literature by highlighting the determinants of early internationalizing firms and their performance with a focus on the entrepreneur, firm, and environment factors. Furthermore, an integrative framework is developed to account for the relationships among determinants, early internationalization, and outcomes. Finally, the authors reveal some significant gaps to advance an important research agenda for future research.

ARTICLE IN PRESS



### 4.9. Culture and international business research: a review and research agenda (Srivastava et al., 2020)

Srivastava, Singh and Dhir (2020) explores the role of culture and international business in internationalization outcomes through a systematic review and analysis of articles published between 2009 and 2019. By mapping the current research domain, their review reflects the avenues for future research in theory development, context, characteristics, and methodology (using TCCM protocol developed by Paul & Rosado-Serrano, 2019). They have identified eight research clusters as follows. (1) national culture, (2) external uncertainty avoidance, (3) knowledge transfer & collaboration, (4) HRM & management practices, (5) international diversification research, (6) entrepreneurial mindset, (7) interaction, and (8) firm performance. The clusters were grouped into independent factors and internationalization outcome factors. Besides, their framework may provide deeper insights into the theoretical implications which will lead to further advancement in these research areas.

## 5. Conclusion

The main purpose of a review article is to critically analyse the extant literature in a given research area, theme or discipline, identifying relevant theories, key constructs, empirical methods, contexts, and remaining research gaps in order to set a future research agenda based on those gaps. We have provided experience-based information in the form of insights and guidelines on how to develop scientifically acceptable and truly impactful literature review articles. It is important to consider these suggestions, at least partly, to avoid rejection of this type of research articles in outstanding business-related journals. These insights are based on our experience as editors of review articles as well as based on the ideas and comments given by a very exclusive group of anonymous reviewers. In our opinion, it is, indeed, an “art” to develop a classic systematic review or a meta-analytical contribution. Although a single author of a highly original contribution has been also included in this Special Review Issue, ideally it seems that a team of two or three scholars are usually required to develop such impactful review articles, so that they can exchange ideas and use the exposure and experience of those who have track record and more accumulated knowledge. We hope readers really enjoy the outcome of our work at least as much as we have enjoyed its development process as guest co-editors.

## Acknowledgements

We fondly acknowledge our greatest recognition to all of the authors who submitted papers for this volume and to the many reviewers who have contributed to reviewing all of them and meaningfully improving especially the articles selected for publication through their diligent effort and devoted time.

## References

Aaltonen, P. H. M. (2020). Piecing together a puzzle—A review and research agenda on internationalization and the promise of exaptation. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2020.101664 in press.

Bahoo, S., Alon, I., & Paltrinieri, A. (2020). Corruption in international business: A review and research agenda. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2019.101660 in press.

Bem, D. J. (1995). Writing a review article for psychological bulletin. *Psychological Bulletin, 118*(2), 172–177.

Callahan, J. L. (2014). *Writing literature reviews: A reprise and update.* https://doi.org/10.1177/1534484314536705.

Canabal, A., & White, G. O., III (2008). Entry mode research: Past and future. *International Business Review, 17*(3), 267–284.

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal. *Academy of Management Journal, 50*(6), 1281–1303.

Coviello, N. E., & Jones, M. V. (2004). Methodological issues in international entrepreneurship research. *Journal of Business Venturing, 19*(4), 485–508.

Dabić, M., Vlačić, B., Paul, J., Dana, L. P., Sahasranamam, S., & Glinka, B. (2020). Immigrant entrepreneurship: A review and research agenda. *Journal of Business Research, 113*, 25–38.

Donnelly, R., & Manolova, T. S. (2020). Foreign location decisions through an institutional lens: A systematic review and future research agenda. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2020.101690 in press.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. *Academy of Management Review, 14*(1), 57–74.

Fetscherin, M., & Heinrich, D. (2015). Consumer brand relationships research: A bibliometric citation meta-analysis. *Journal of Business Research, 68*(2), 380–390. https://doi.org/10.1016/j.jbusres.2014.06.010.

Fisch, C., & Block, J. (2018). Six tips for your (systematic) literature review in business and management research. *Management Review Quarterly, 68*, 103–106. https://doi.org/10.1007/s11301-018-0142-x.

Furrer, O., Thomas, H., & Goussevskaia, A. (2008). The structure and evolution of the strategic management field: A content analysis of 26 years of strategic management research. *International Journal of Management Reviews, 10*(1), 1–23.

Gilal, F. G., Zhang, J., Paul, J., & Gilal, N. G. (2019). The role of self-determination theory in marketing science: An integrative review and agenda for research. *European Management Journal, 37*(1), 29–44. https://doi.org/10.1016/j.emj.2018.10.004.

Gupta, P., Chauhan, S., Paul, J., & Jaiswal, M. P. (2020). Social entrepreneurship research: A review and future research agenda. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2020.03.032.

Hao, A. W., Paul, J., Trott, S., Guo, C., & Wu, H. H. (2019). Two decades of research on nation branding: A review and future research agenda. *International Marketing Review*. https://doi.org/10.1108/IMR-01-2019-0028.

Hoskisson, R. E., Wan, W. P., Yiu, D., & Hitt, M. A. (1999). Theory and research in strategic management: Swings of a pendulum. *Journal of Management, 25*(3), 417–456.

Hulland, J., & Houston, M. B. (2020). Why systematic review papers and meta-analyses matter: An introduction to the special issue on generalizations in marketing. *Journal of the Academy of Marketing Science, 48*, 351–359. https://doi.org/10.1007/s11747-020-00721-7.

İpek, İ., & Bıçakcıoğlu-Peynirci, N. (2019). Export market orientation: An integrative review and directions for future research. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2019.101659 101659.

Ji, J., Plakoyiannaki, E., Dimitratos, P., & Chen, S. (2019). The qualitative case research in international entrepreneurship: A state of the art and analysis. *International Marketing Review, 36*(1), 164–187. https://doi.org/10.1108/IMR-02-2017-0052.

Jiang, G., Kotabe, M., Zhang, F., Hao, A. W., & Wang, C. L. (2020). The determinants and performance of early internationalizing firms: A literature review and research agenda. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2019.101662 in press.

Jones, M. V., Coviello, N., & Tang, Y. K. (2011). International entrepreneurship research (1989–2009): A domain ontology and thematic analysis. *Journal of Business Venturing, 26*(6), 632–659.

Kahiya, E. T. (2018). Five decades of research on export barriers: Review and future directions. *International Business Review, 27*(6), 1172–1188. https://www.sciencedirect.com/science/article/abs/pii/S0969593117303141.

Klier, H., Schwens, C., Zapkau, F. B., & Dikova, D. (2017). Which resources matter how and where? A meta-analysis on firms' foreign establishment mode choice. *Journal of Management Studies, 54*(3), 304–339. https://doi.org/10.1111/joms.12220.

Knoll, J., & Matthes, J. (2017). The effectiveness of celebrity endorsements: A meta-analysis. *Journal of the Academy of Marketing Science, 45*(1), 55–75.

Kozlenkova, I. V., Samaha, S. A., & Palmatier, R. W. (2014). Resource-based theory in marketing. *Journal of the Academy of Marketing Science, 42*(1), 1–21. https://doi.org/10.1007/s11747-013-0336-7.

Kraus, S., Breier, M., & Dasí-Rodríguez, S. (2020). The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal,* 1–20. https://doi.org/10.1007/s11365-020-00635-4.

Kumar, A., Paul, J., & Unnithan, A. B. (2019). 'Masstige' marketing: A review, synthesis and research agenda. *Journal of Business Research, 113*, 384–398. https://doi.org/10.1016/j.jbusres.2019.09.030 May 2020.

Marabelli, M., & Newell, S. (2014). Knowing, power and materiality: A critical review and reconceptualization of absorptive capacity. *International Journal of Management Reviews, 16*(4), 479–499.

Niittymies, A., & Pajunen, K. (2020). Cognitive foundations of firm internationalization: A systematic review and agenda for future research. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2019.101654 in press.

Palmatier, R. W., Houston, M. B., & Hulland, J. (2018). Review articles: Purpose, process, and structure. *Journal of Academy of Marketing Science*. https://doi.org/10.1007/s11747-017-0563-4.

Pati, D., & Lorusso, L. N. (2018). How to write a systematic review of the literature. *HERD: Health Environments Research & Design Journal, 11*(1), 15–30.

Paul, J. (2019). Marketing in emerging markets: A review, theoretical synthesis and extension. *International Journal of Emerging Markets, 15*(3), 446–468. https://doi.org/10.1108/IJOEM-04-2017-0130.

Paul, J., & Benito, G. R. (2018). A review of research on outward foreign direct investment from emerging countries, including China: What do we know, how do we know and where should we be heading? *Asia Pacific Business Review, 24*(1), 90–115. https://doi.org/10.1080/13602381.2017.1357316.

Paul, J., & Feliciano-Cestero, M. (2020). Five decades of research on foreign direct investment by MNEs: An overview and research agenda. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2020.04.017.

Paul, J., & Mas, E. (2019). Toward a 7-P framework for international marketing. *Journal of Strategic Marketing,* 1–21. https://doi.org/10.1080/0965254X.2019.1569111.

Paul, J., & Rosado-Serrano, A. (2019). Gradual internationalization vs born-global/

ARTICLE IN PRESS



international new venture models: A review and research agenda. *International Marketing Review, 36*(6), 830–858. https://doi.org/10.1108/IMR-10-2018-0280.
Paul, J., & Singh, G. (2017). The 45 years of foreign direct investment research: Approaches, advances and analytical areas. *The World Economy, 40*(11), 2512–2527. https://doi.org/10.1111/twec.12502.
Paul, J., Parthasarathy, S., & Gupta, P. (2017). Exporting challenges of SMEs: A review and future research agenda. *Journal of World Business, 52*(3), 327–342. https://doi.org/10.1016/j.jwb.2017.01.003.
Piper, R. J. (2013). How to write a systematic literature review: A guide for medical students. *National AMR, Fostering Medical Research,* 1–8.
Post, C., Sarala, R., Gatrell, C., & Prescott, J. E. (2020). Advancing theory with review articles. *Journal of Management Studies, 57*(2), https://doi.org/10.1111/joms.12549.
Randhawa, K., Wilden, R., & Hohberger, J. (2016). A bibliometric review of open innovation: Setting a research agenda. *Journal of Product Innovation Management, 33*(6), 750–772. https://doi.org/10.1111/jpim.12312.
Rauch, A., Wiklund, J., Lumpkin, G. T., & Frese, M. (2009). Entrepreneurial orientation and business performance: An assessment of past research and suggestions for the future. *Entrepreneurship Theory and Practice, 33*(3), 761–787.
Reuber, A. R. (2010). Strengthening your literature review. *Family Business Review, 23*(2), 105–198.
Rialp, A., Rialp, J., & Knight, G. A. (2005). The phenomenon of early internationalizing firms: What do we know after a decade (1993–2003) of scientific inquiry? *International Business Review, 14*(2), 147–166.
Rialp, A., Rialp, J., & Knight, G. A. (2014). *International entrepreneurship: A review and future directions. The Routledge companion to international entrepreneurship.* Routledge27–48.
Rialp, A., Merigó, J. M., Cancino, C. A., & Urbano, D. (2019). Twenty-five years (1992–2016) of the International Business Review: A bibliometric overview. *International Business Review, 28*(6), https://doi.org/10.1016/j.ibusrev.2019.101587.
Rindfleisch, A., & Heide, J. B. (1997). Transaction cost analysis: Past, present, and future applications. *Journal of Marketing, 61*(4), 30–54. https://doi.org/10.2307/1252085.
Rosado-Serrano, A., Paul, J., & Dikova, D. (2018). International franchising: A literature review and research agenda. *Journal of Business Research, 85*, 238–257. https://doi.org/10.1016/j.jbusres.2017.12.049.
Schmid, S., & Kotulla, T. (2011). 50 years of research on international standardization and adaptation—From a systematic literature analysis to a theoretical framework. *International Business Review, 20*(5), 491–507.
Schmid, D., & Morschett, D. (2020). Decades of research on foreign subsidiary divestment: What do we really know about its antecedents? *International Business Review*. https://doi.org/10.1016/j.ibusrev.2019.101653 in press.
Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339.
Sorescu, A., Warren, N. L., & Ertekin, L. (2017). Event study methodology in the marketing literature: An overview. *Journal of the Academy of Marketing Science, 45*(2), 186–207. https://doi.org/10.1007/s11747-017-0516-y.
Srivastava, S., Singh, S., & Dhir, S. (2020). Culture and international business research: A review and research agenda. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2020.101709.
Tang, R. W., & Buckley, P. J. (2020). Host country risk and foreign ownership strategy: Meta-analysis and theory on the moderating role of home country institutions. *International Business Review*. https://doi.org/10.1016/j.ibusrev.2020.101666 in press.
Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. *Human Resource Development Review, 15*(4), 404–428. https://doi.org/10.1177/1534484316671606.
Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: An analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science, 44*(1), 119–134. https://doi.org/10.1007/s11747-015-0455-4.
Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), 13–23.
Xie, E., Reddy, K. S., & Liang, J. (2017). Country-specific determinants of cross-border mergers and acquisitions: A comprehensive review and future research directions. *Journal of World Business, 52*(2), 127–183. https://doi.org/10.1016/j.jwb.2016.12.005.

**Dr Justin Paul**, serves as Editor-in-chief of International Journal of Consumer studies (IJCS), and a full professor, University of Puerto Rico, USA and a 'Distinguished Scholar' with IIM- K, India's premier business school. A former faculty member with the University of Washington, he is known as an author/co-author of books such as Business Environment (4th ed), International Marketing, Services Marketing, Export-Import Management (2nd edition) by McGraw-Hill & Oxford University Press respectively. He has served as Lead Guest editor for Journal of Business Research, Journal of Retailing & Consumer Services, Small Bus Economics and European Business Review. Dr. Paul introduced Masstige model and measure for brand management, CPP Model for internationalization, SCOPE framework for Small firms and 7-P Framework for International Marketing. His articles have been downloaded over 600,000 times during last five years. He has published over 60 research papers in SSCI journals and 75 in Scopus. He has also served as a faculty member of Nagoya University, Japan and IIM. In addition, he has taught full courses at Aarhus University- Denmark, Grenoble Eco le de Management-& Universite De Versailles -France, University-Lithuania, Warsaw -Poland and has conducted research development workshops in countries such as Austria, USA, Spain, Croatia, China.

**Alex Rialp-Criado** is Associate professor at Universitat Autònoma de Barcelona (UAB), Spain, and Adjunct Professor at the Norwegian University of Science and Technology (NTNU). His research covers international business/marketing and international entrepreneurship domains, with a focus on the internationalization of new ventures and established SMEs. He is author or co-author, and guest editor of different books, chapter books, and articles in leading international academic journals in these fields such as: International Business Review, Management International Review, Journal of World Business, Journal of International Marketing, International Marketing Review, Journal of Small Business Management, Entrepreneurship and Regional Development, European Management Journal, Journal of International Entrepreneurship, Critical Perspectives on International Business, Journal of Global Information Management, and Advances in International Marketing, among others. Dr Alex Rialp also serves as editorial board member of the International Business Review and the Journal of International Entrepreneurship, as well as ad-hoc reviewer for many different academic journals.

Quality & Quantity
https://doi.org/10.1007/s11135-020-01059-6

# The ABC of systematic literature review: the basic methodological guidance for beginners

**Hayrol Azril Mohamed Shaffril[1] · Samsul Farid Samsuddin[2] · Asnarulkhadi Abu Samah[1,3]**

Accepted: 12 October 2020
© Springer Nature B.V. 2020

**Abstract**
There is a need for more methodological-based articles on systematic literature review (SLR) for non-health researchers to address issues related to the lack of methodological references in SLR and less suitability of existing methodological guidance. With that, this study presented a beginner's guide to basic methodological guides and key points to perform SLR, especially for those from non-health related background. For that, a total of 75 articles that passed the minimum quality were retrieved using systematic searching strategies. Seven main points of SLR were discussed, namely (1) the development and validation of the review protocol/publication standard/reporting standard/guidelines, (2) the formulation of research questions, (3) systematic searching strategies, (4) quality appraisal, (5) data extraction, (6) data synthesis, and (7) data demonstration.

**Keywords** Systematic literature review · Basic methodology · Non-health related studies · Beginners

## 1 Introduction

Previous works are fundamental to the creation of new knowledge. When performing literature review, researchers analyse, interpret, and critically evaluate the existing body of knowledge. The process allows them to discover the patterns of prior results, comprehend the depth and details of the existing knowledge, and identify gaps for further exploration.

✉ Hayrol Azril Mohamed Shaffril
hayrol82@gmail.com

Samsul Farid Samsuddin
samsulfarid@gmail.com

Asnarulkhadi Abu Samah
asnarulhadi@gmail.com

[1] Institute for Social Science Studies, Universiti Putra Malaysia, Putra Infoport, 43400 Serdang, Selangor, Malaysia

[2] Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[3] Faculty of Human Ecology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Published online: 23 October 2020

Traditional literature review has been long practised and regarded as one of the best ways to situate a study within the existing knowledge. Currently, there is another alternative of a review, which is systematic literature review (SLR). According to Higgins et al. (2011), SLR or also known as systematic review (SR) can be defined as follows:

> A systematic literature review aims to comprehensively locate and synthesise related research using organised, transparent, and replicable procedures at each step in the process.

SLR has several advantages compared to traditional review such as its numerous unique procedures. SLR encourages researchers to look for studies outside their own subject areas and networks through the introduction of extensive searching methods, predefined search strings, and standard inclusion and exclusion criteria (Robinson and Lowe 2015). This kind of review stress on transparency, all terms in inclusion criteria for example, must be defined and justified while exclusion of articles must be reasoned (Greyson et al. 2019). Furthermore, SLR heavily focuses on evidence, impact, validity and causality, it urges researchers to examine information on research design, analytical methods and causal chains, and by practising this, SLR is controlling the quality of review by ensuring the robustness of evidence (Lockwood et al., 2015; Mallet et al. 2012). The use of review protocol, publication standard or established guidelines on the other hand is effective in guiding and ensuring the researcher are 'on track' and also improving the methodological transparency of the review (Haddaway et al. 2018). Without SLR, the review of existing literature is less comprehensive and the extent of comprehensiveness of literature review cannot be answered; furthermore, transparency can be a serious issue in several types of review without SLR (Dixon-wood et al. 2005).

The preference of systematic reviews in health-related and other sciences-related studies is different from other fields of studies as they have complicated research questions, various theoretical and epistemological methodologies, and abundant sources of information (Berrang-Ford et al. 2015). There are major differences in the development of SLR particularly in terms of methodology in other field of studies. Berrang-Ford et al. (2015), stressed that such situation, if unintentional, is caused by monopoly of the use and application of systematic reviews within health and associated science. Berrang-Ford et al. (2015) further explained the domination of health related studies when they demonstrated significant difference in terms of number of review articles, as they found only 1% paper from non-health fields with terms of systematic review in Web of Science database.

Lacking of methodological guidance of non-health related in SLR can be seen in several aspects. In terms of review, protocol for example, those from non-health need to rely or adapt on health-based review protocol (e.g. Cochrane, Campbell, Joanna-Briggs) as there are none to few guidance from non-health related that are available for them. As they are from different field of studies, relying on or adapting these review protocols are making SLR of non-health related to fail to fully abide by standard guidance which leads to limited penetration and use of systematic approaches (Haddaway et al. 2018).

Furthermore, for those who interested to use the publication standard of Preferred reporting items for systematic reviews and meta-analyses (PRISMA) albeit claimed by its inventor, Moher et al. (2009), can either be used for a systematic review that focuses on randomised trials or as a basis for other types of research particularly intervention, is actually facing several issues especially when it comes to reviewing involving qualitative and mixed-method research designs (Haddaway et al. 2018). Therefore, issues related to lacking of methodological guidance and the less suitability of existing methodological

references have led to the need for more methodological-based articles on SLR for non-health researchers.

### 1.1 Aim and objective

The present article aims to solve the issue related to lacking of methodological guidance and less suitability of existing methodological references by providing a general understanding of basic methodologies for SLR that can guide non-health related scholars in their review. The guidance is based on seven main aspects of SLR methodology as follows: (1) the development and validation of the review protocol/publication standard/reporting standard/guidelines; (2) the formulation of research questions; (3) systematic searching strategies; (4) quality appraisal; (5) data extraction; (6) data synthesis; and (7) data demonstration. The present article is based on the review related to the methodology that has been practised by previous scholars in their SLR. It provides diverse perspectives and practices by previous scholars that offer more options, guidance, ideas, and understanding of SLR for beginners.

Although there are several SLR guidelines available for researchers such (see Xiao and Watson, 2019), the current study presented a unique set of SLR guidelines. The main difference between the current study and the study by Xiao and Watson (2019) lies in how SLR is initiated - the current study emphasised on starting SLR by developing a review protocol or referring to the available guidelines (e.g. publication standards or other established guidelines), rather than formulating research question. This proposed set of SLR guidelines was established based on several justifications. Firstly, developing a specific review protocol or referring to other established guidelines can assist researchers to plan on the key aspects of their research and include or write in their SLR, which eventually allow them to present a transparent, transferable and replicable (Mengist et al. 2020). Secondly, developing a specific review protocol or referring to other established guidelines as a first step in SLR can assist researchers to formulate good and comprehensive research questions, strategise their systematic searching efforts, select suitable inclusion criteria, have a rigorous process of quality appraisal, strategise their data extraction process and data synthesis, and demontsrate the best data of their review (del Amo et al. 2018).

Furthermore, the current study was expected to advance the current understanding of SLR guidance by providing insights from multifaceted perspectives to the suggested methodology of Xiao and Watson (2019). The current study discussed different methodological perspectives with regards to searching functions - for example, Xiao and Watson (2019) focused on Boolean operator, while this study presented more searching functions such as phrase searching, truncation, wildcard and field code functions (please see Table 2). Although Xiao and Watson (2019) stressed the importance of validating the developed review protocol in planning specific methodology for the review, the current study demonstrated the capacity of publication or reporting standard or other established guidelines to perform a similar purpose. As for the process of formulating research question, the current study and Xiao and Watson (2019) shared a similar view on the importance of having general or specific research questions. However, the current study viewed certain aspects differently - the proposed guidelines include informing the differences between quantitative and qualitative research questions and explaining the ability of research questions development tools such as PICO, PICOS, SPIDER and TOPICS + M, in assisting researchers to formulate the best research questions for their review. Meanwhile, as for the scope of database selection, Xiao and Watson (2019) discussed some of the commonly

used databases such as Web of Science, EBSCO, ProQuest, IEEE Xplore, Google Scholar, DOAJ and OAIster. However, the current study demonstrated the importance of 14 databases in SLR, in which Gusunbauer and Haddaway (2020) proved its strength in terms of coverage and searching abilities. The current study also demonstrate the importance of using Google Scholar as a supporting database and not as a leading database in systematic searching strategy. Numerous comparisons can be made between the current study and the study by Xiao and Watson (2019). Nevertheless, the discussed differences in this study were expected to advance the understanding and enrich the existing literature from multifaceted perspectives, which allow researchers to select the best methods for their SLR.

## 2 Methodology

### 2.1 The process of retrieving articles for the review

The authors divided the searching process based on seven objectives to retrieve relevant articles. The appropriate keywords were selected based on each objective for the searching process (refer to Table 1). This study used two main searching techniques, namely advanced searching (on the selected databases) and manual searching on four main databases which were Scopus, Science Direct, Google Scholars, and Google engine search. The authors also used the phrase searching function and the Boolean operator OR or/and AND to combine keywords in their advance searching process. This study used three main techniques for manual searching, namely handpicking, backward tracking, and forward tracking.

The first stage of keywords searching obtained 141 potential articles. Then, the authors determined the inclusion criteria as follows: content of the selected articles, timeline publication and language. As the paper are much related to guidance on developing an SLR, the content of the selected articles must be heavily focus on SLR related methodology. Without denying the importance of publication before 2005, the authors are relying on Okoli and Schabram (2010) who stress on the impossibility of reviewing all the published articles in the entire human history and decided to choose a timeline between 2005-2020 as it is in line with the concept of study's maturity by Kraus et al. (2020), whereby within this period, number of relevant articles is higher and therefore more major topics are investigated and more evidence driven. Furthermore, during this timeline publication, important articles such as by Dixon Wood et al. (2005), Petticrew and Roberts (2006) and Whitemore and Knafl (2005) have advanced important methodological concepts of SLR (e.g. integrative review; mixed-method synthesis) and should be considered by the authors. Authors have decided to review only articles published in English, as Linares-Espinos et al. (2018) stressed the importance of selecting publications in languages that they understand as articles in foreign languages can create further confusion, add more costs for the review, and consume time.

Next, two reviewers independently screened relevant articles that match these criteria by examining titles and abstracts as well as results and methodology sections. This study only selected articles that are mutually agreed by both reviewers and any disagreement on articles selection was resolved by discussions. This process has excluded 66 articles which left 75 articles for quality assessment.

**Table 1** Articles included in the review

| | Main keywords used | Number of resulted articles to be considered in the review | Number of articles included in the review |
|---|---|---|---|
| Review protocol/publication standard/reporting standard/guidelines etc. | Review Protocol<br>Publication Standards<br>Reporting Standards<br>Guidelines<br>Systematic literature review<br>Systematic review<br>Cochrane<br>Campbell<br>PRISMA<br>RAMESES | 20 | *10 articles reviewed*<br>1. Page et al. (2020)<br>2. Kushwah et al. (2019)<br>3. Reim et al. (2015)<br>4. Moher et al. (2009)<br>5. Liberati et al. (2009)<br>6. Wong et al. (2013)<br>7. Haddaway et al. (2018)<br>8. Petticrew and Roberts (2006)<br>9. Kitchenham and Charters (2007)<br>10. Durach et al. (2017) |
| Formulation of research question | Research questions<br>Systematic literature review<br>Systematic review | 25 | *15 articles reviewed*<br>1. Burgers et al. (2019)<br>2. Johnson and Hennessy (2019)<br>3. Okoli (2015)<br>4. Xiao and Watson (2019)<br>5. Thomas et al. (2020)<br>6. Petticrew and Roberts (2006)<br>7. Doody and Bailey (2016),<br>8. Onwuegbuzie and Leech (2007)<br>9. Schardt et al. (2007).<br>10. Palaskar (2017)<br>11. Cañón and Buitrago-Gómez (2018)<br>12. Centre for Reviews and Dissemination (2006)<br>13. Cooke et al. (2012)<br>14. Mantzoukas (2008)<br>15. Creswell (2013) |

**Table 1** (continued)

| | | Main keywords used | Number of resulted articles to be considered in the review | Number of articles included in the review |
|---|---|---|---|---|
| Systematic searching strategy | Identification | Keywords<br>Formulation of keywords<br>Systematic literature review<br>Systematic review | 30 | *19 articles reviewed*<br>1. Wanden-Berghe and Sanz-Valero (2012)<br>2. Kitchenham and Charters (2007)<br>3. Methley et al. (2014)<br>4. Petticrew and Roberts (2006)<br>5. Bates et al. (2017)<br>6. Xiao and Watson (2019)<br>7. Younger (2010)<br>8. Gusenbauer and Haddaway (2020)<br>9. Cooper et al. (2018)<br>10. Athukorala et al. (2016)<br>11. Gusenbauer (2019)<br>12. Housyar and Sotudeh (2018)<br>13. Halevi et al. (2017)<br>14. Fagan (2017)<br>15. Haddaway et al. (2015)<br>16. Shaffril et al. (2018)<br>17. Thomas et al. (2017)<br>18. Levy and Ellis (2006)<br>19. Kastner et al. (2007) |
| | Screening | Inclusion criteria<br>Exclusion criteria<br>Inclusion and exclusion criteria<br>Systematic literature review<br>Systematic review | 12 | *9 articles reviewed*<br>1. Patino and Ferreira (2018)<br>2. Xiao and Watson (2019)<br>3. Okoli (2015)<br>4. Johnson and Hennessy (2019)<br>4. Kitchenham and Charters (2007)<br>5. Linares-Espinos et al. (2018)<br>6. Delaney and Tamas (2018)<br>7. Okoli and Schabram (2010)<br>8. Kraus et al. (2020) |

**Table 1** (continued)

| | | Main keywords used | Number of resulted articles to be considered in the review | Number of articles included in the review |
|---|---|---|---|---|
| | Eligibility | Manual screening<br>Eligibility<br>Systematic literature review<br>Systematic review | 4 | *2 articles reviewed*<br>1. Moher et al. (2009)<br>2. Liberati et al. (2009) |
| Quality appraisal | | Quality appraisal<br>Appraise of quality<br>Bias<br>Systematic literature review<br>Systematic review | 15 | *9 articles reviewed*<br>1. Higgins et al. (2019)<br>2. Seehra et al. (2016)<br>3. Siering et al. (2013)<br>4. Hannes (2011)<br>5. Pace et al. (2012)<br>6. Long and Godfrey (2004)<br>7. Hong et al. (2018)<br>8. Charrois (2015)<br>9. Petticrew and Roberts (2006) |
| Data extraction | | Data extraction<br>Extracting data<br>Systematic literature review<br>Systematic review | 5 | *3 articles reviewed*<br>1. Gomersall et al. (2015)<br>2. Charrois (2015)<br>3. Kitchenham and Charters (2007) |

**Table 1** (continued)

| | Main keywords used | Number of resulted articles to be considered in the review | Number of articles included in the review |
|---|---|---|---|
| Data analysis | Data analysis<br>Data synthesis<br>Quantitative analysis<br>Quantitative synthesis<br>Qualitative synthesis<br>Qualitative analysis<br>Mixed method synthesis<br>Mixed method analysis<br>quantitative and qualitative analysis<br>Quantitative and qualitative synthesis<br>Systematic literature review<br>Systematic review | 23 | *20 articles reviewed*<br>1. Morton et al. (2018)<br>2. Shorten and Shorten (2013)<br>3. Rousseau et al. (2008)<br>4. Green et al. (2006)<br>5. Braun and Clark (2006)<br>6. Noble and Mitchell (2016)<br>7. Patterson (2012)<br>8. Patterson et al. (2001)<br>9. Wong et al. (2013)<br>10. Brunton et al. (2006)<br>11. Dixon-wood et al. (2005)<br>12. Sandelowski et al. (2007)<br>13. Soares et al. (2013)<br>14. Hopia et al. (2016)<br>15. Prieto and Rumbo-Prieto (2018)<br>16. Sandelowski et al. (2006)<br>17. Mays et al. (2005)<br>18. Barnett-Page and Thomas (2009)<br>19. Flemming et al. (2018)<br>20. Whitemore and Knafl (2005) |
| Demonstration of findings | Reporting<br>Data demonstration<br>Systematic literature review<br>Systematic review | 6 | *4 articles reviewed*<br>1. Moher et al. (2009)<br>2. Shaffril et al. (2019)<br>3. Peters et al. (2015)<br>4. Petticrew and Roberts (2006) |

### 2.2 Quality assessment of the selected articles/documents

The qualities of the remaining articles were independently assessed by two reviewers that focus on abstract, method, and main results. Petticrew and Roberts (2006) suggest that reviewers should qualitatively assess the articles by categorising them into low, moderate, or high level of quality. Articles only be included if it meets high or moderate quality. Adapting on Hong et al. (2018) guidance, the reviewers are guided by these five criteria on their quality assessment namely (1) Are the articles' main aim related to SLR methodology?; (2) Are the articles providing all methodology needed in developing SLR?; (3) Are the articles clearly defined the SLR methodology to the authors?; (4) Is there an adequate rationale for each of the guidance on SLR methodology in the articles?; and (5) Are the articles providing any option or alternative to their suggested guidance on SLR methodology? For each of the criteria, the reviewers have three options of answer namely yes, no or can't tell. The reviewers decided that if the articles fulfill four or five criteria, then the articles are in a high level of quality, if the articles fulfill at least three criteria, then the categorized it as moderate in quality and if the articles fulfill merely one or two criteria, then the articles are in a low level quality. The reviewers of this study mutually agreed that 69 articles met the minimum requirement (high or moderate). They further discussed the suitability of the remaining six articles in the review. Although these six articles are categorized as low, the reviewers believed that there are some SLR related methods implemented in the articles, can be recommended to other SLR authors. Shaffril et al. (2019) for example, although is not a methodology paper, however, have practiced modification of flow diagram to fit their SLR and this kind of practice should be considered by others. Due to this situation, the reviewers have decided to maintain these six articles in the review. The final amount of selected articles was 75.

This study only extracted data from selected studies that fit the objectives. The process was performed independently by all the authors. The co-authors provided advice to the lead authors when finalising the suitability of the extracted data. The study used thematic synthesis to analyse the data by deductively developing the theme based on seven aspects of SLR as mentioned earlier.

## 3 Results and discussions

This section discusses the answers to the research questions. It discusses the following: (1) development and validation of review protocol which are guided by publication standards or reporting standards or established guidelines; (2) formulating the research question; (3) systematic searching strategies; (4) quality appraisal; (5) data extraction; (6) data synthesis; and (7) demonstration of findings.

## 3.1 The development and validation of the review protocol/publication standard/reporting standard/guidelines

### 3.1.1 Development and validation of review protocol

The first step in SLR is developing and validating the review protocol. The review protocol is the researchers' plan on things that they want to consider and include in their review. The protocol describes the reasons for the review question and proposed methods besides reducing bias and explaining how the different types of study will be positioned, evaluated, and produced. It also includes details related to the review's dissemination strategy. It is common to alter the protocols of systematic review after the review such as the methodology section. Page et al. (2020) stated the need to provide justification and document the changes. Researches can refer to several established review protocols, including BEME Collaboration, Joanna Briggs Institute, Campbell Collaboration, and Cochrane Collaboration. The use of multiple review protocols is advantageous for rigorous systematic review, bias reduction in data selection and analysis besides paving ways for others to adopt a similar protocol for verification and cross-checking stages (Page et al. 2020).

### 3.1.2 Referring to a publication/reporting standard

Some non-health studies that practised experimental based research can refer to the review protocol at a general or specific scale. However, things are not the same to certain field of studies such as art and humanities, engineering, computer science, and architecture. Notably, review protocol is not the only guide for SLR as some researchers have developed publication standards/reporting standards that enable researchers to assess the reviews' quality and accuracy. Moher et al. (2009) and Liberati et al. (2009) have developed Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) that suggests the lowest number of items in systematic reviews and meta-analyses reporting. PRISMA has 27 items for the review. The main priority of PRISMA is randomised trials. However, it can be the guide for the systematic reviews of other fields of study which involve the assessments of interventions. Wong et al. (2013) developed Realist and MEta-narrative Evidence Syntheses: Evolving Standards (RAMESES), which is a publication standard for realist syntheses. RAMESES has 19 items for the review.

Furthermore, scholars had developed a specific publication or reporting standard for their respective field of studies due to these two main reasons: (1) SLR is getting more attention from non-health related studies; and (2) the inability of non-health studies to refer or adapt to available review protocol or publication/reporting standards which are specifically tailored for health-related studies. For example, Haddaway et al. (2018) created Reporting standards for Systematic Evidence Syntheses: pro forma (ROSES: pro forma), which is a reporting standard for environmental-related studies. Haddaway et al. (2018) listed nine main differences between PRISMA and ROSES, which include the following: ROSES is specifically created for systematic maps and environmental systematic reviews, it did not emphasise on quantitative synthesis, and it accommodates other types of synthesis.

### 3.1.3 Referring to established guidelines

Besides that, several scholars have developed established guidelines in developing systematic literature review for their respective field of studies such as software engineering (Kitchenham and Charters 2007), social science (Petticrew and Roberts 2006), and supply chain management (Durach et al. 2017). SLR has the concept of replicate (Higgins et al. 2011). Reim et al. (2015) and Kushwah et al. (2019) refer to previous studies as a guide for SLR.

The following are the main key points of first step of SLR

- Any SLR must be guided by at least review protocol/publication standard/reporting standard/established guidelines/past published SLR articles.
- Most of the review protocol/publication standards are tailored for health-related studies.

## 3.2 The formulation of research questions

### 3.2.1 General and specific research questions

A research question must guide SLR that drives the entire process. The methodology, extracted and synthesised data must be able to answer the research questions. Some researchers claimed that research questions should not be too general that might result in more selected articles, time consumption, and difficulty to compare and manage data (Burgers et al. 2019; Johnson and Hennessy 2019). Some researchers suggested the following: formulating a specific research question (Okoli 2015) and selecting a subtopic for the review (Xiao and Watson 2019). Petticrew and Roberts (2006) affirm that research questions must not be too specific as there might be too few articles and SLR cannot be conducted on a small sample. On the other hand, Thomas et al. (2020) claimed that broad research questions could offer a wide-ranging summary of previous findings to explore the consistency of results which can lead to generalisability.

### 3.2.2 Quantitative research questions

There is a difference between the formulation of quantitative and qualitative research questions. Doody and Bailey (2016) explain that the formulation of quantitative research questions can be in three forms, namely descriptive, comparative, and relationship. They mention that researchers should avoid words like ‘do’, ‘does’, ‘is’, and ‘are’ that can trigger yes or no responses. Onwuegbuzie and Leech (2007) describe the three types of quantitative research questions namely descriptive, comparative and relationship. PICO model is commonly used to formulate the clinical question and also research question which assists the researchers in data extraction processes and finding relevant evidences from database. PICO has four main criteria, namely Population, Intervention, Control, and Outcomes (Schardt et al. 2007). Palaskar (2017) revealed several PICO advantages such as the improvement of precision and conceptual clarity of clinical problems by frames in which information are offered in pre-search reference interviews that lead to more complicated search approaches for accurate search findings. Canon and Butrigo-Gomes (2018) criticised PICO for its difficulties to prioritise questions, the lack of time to formulate and answer them, the lack of a tool to review literature effectively, besides being too rigid and unsuitable for the current research setting. Besides PICO, there are others models or tools

that can assist researchers to develop their research questions. Centre for Reviews and Dissemination (2006) and Cooke et al. (2012) for example have invented population, intervention, control, outcomes and study design (PICOS) and sample, the phenomenon of interest, design, evaluation and research type (SPIDER) as alternatives. Johnson and Hennessy (2019) on the other hand introduced Time, Outcome, Population, Intervention, Comparison, and Study Designs (TOPICS + M) in helping formulating research questions.

.

### 3.2.3 Qualitative research questions

Qualitative research questions are more flexible, adaptable, and have no directions that focus on either general or specific areas of research or sub-categories (Creswell 2013). Sub-questions can solve problems and difficulties of explaining the content that aimed to determine or explore a process or define experiences. Mantzoukas (2008) noted that researchers should scrutinise the content of their research, the link between the content with the theoretical plans, and the structure.

Among key points that should be considered in the stage of formulating the research questions are.

- Research question that guides the entire SLR process.
- Researchers should be able to differentiate between quantitative and qualitative research questions.
- PICO, PICOS, and SPIDER can assist researchers in formulating appropriate research questions for their SLR.

## 3.3 Systematic searching strategies

This section describes the three sub-processes of systematic searching strategies, namely identification, screening, and eligibility.

### 3.3.1 Identification

**3.3.1.1 Selection and enriching the selected keywords** Researchers enrich basic keywords during identification. When researchers use more keywords, the database can retrieve more potential articles. Before determining the appropriate keywords, several basic concepts need to be understood. According to Wanden-Berghe and Sanz-Valero (2012), researchers should put equal focus on comprehensiveness and accuracy when searching. More general keywords will produce more articles, but it might include too many irrelevant articles. On the other hand, too specific keywords will result in more relevant articles but there is the risk of losing records.

Kitchenham and Charters (2007) state that basic keywords can be obtained from research questions and researchers can categorise research questions into specific domains. For instance, researchers can create domains such as climate change, adaptation, and Asian fishermen for the research question, what are the climate change adaptation strategies practised by the Asian fishermen? These three keywords are inadequate and need to pass the identification process to enrich the keywords by identifying their synonyms, related terms, and variation. During the process, the researcher will discover that they can also use keywords like global warming, extreme weathers, La-Nina, and El-Nino. Instead of using

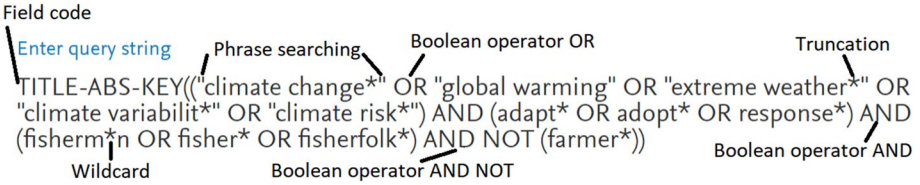


**Fig. 1** An example of symbols and coding in a search/query string developed in Scopus

adaptation, they can use adoption or responses. Other alternatives of Asian fishermen are Asian fisherfolks and Asian fishers.

The identification process that identifies synonyms, related terms, and variation can use some sources such as online thesaurus. However, researchers should carefully select the suggested synonyms as not all the suggested terms are suitable. They can also refer to the keywords used by past studies. It is common for researchers to refer to previous studies in gaining some ideas such as keywords used by previous studies. Besides that, they can use the keywords suggested by the database. Indexing databases such as Scopus and Web of Science can maximise the function through the keywords that have been used by previous studies as researchers can select appropriate keywords for their study. They can also refer to experts' opinions when assessing and validating the comprehensiveness of the search (Petticrew and Roberts 2006).

The keywords with synonyms, related terms, and variation can be useful when searching for suitable articles in the database. The searching process can be done by either typing the keywords in the database engine or using the advanced search function in a certain database. Researchers can use search/query string in the database (see Fig. 1), it is a combination of symbols and coding that allow them to combine all the keywords when searching and avoid repeated searching. Researchers should know the basic symbols and coding such as Boolean operator, phrase searching, truncation, wildcard, and field code functions (see Table 2) to perform search/query string. Researchers can use the online learning platform to learn some indexing databases engines such as Scopus and Web of Sciences.

PICO, PICOS, and SPIDER have been used to detect appropriate keywords in the searching process besides formulating research questions. Methley et al. (2014) suggested that SPIDER and PICOS are useful in identifying suitable keywords for the qualitative systematic review. They also suggested that researchers should use PICO for various databases.

**3.3.1.2 Selection of databases** Researchers are concerned with the quantity and types of databases that should be used in SLR. It is believed that there is no perfect database. Although some databases have superior functions and advantages, they still have weaknesses such as low sensitivity towards keywords and limited searching functions (Bates et al. 2017). Xiao and Watson (2019) revealed that no database is comprehensive and the systematic search must be made from several databases. Younger (2010) claimed that this technique would allow the selected databases to complement their weaknesses.

Gusenbauer and Haddaway (2020) suggested the use of 14 databases that have the strength to act as the leading database in article searching process as follows: ACM Digital Library, BASE, ClinicalTri-als.gov, Cochrane Library, EbscoHost (tested for ERIC, Medline, EconLit, CINHAL Plus, SportsDis-cus), OVID (tested for Embase, Embase Classic, PsychINFO), ProQuest (tested for Nursing and Allied Health Database, Public Health

**Table 2** Selected symbols and coding used in search/query string process in Scopus database

| Functions | Symbols/coding | Purpose |
|---|---|---|
| Field codes | | To search part of documents, more than 60 field codes available in Scopus. Four of them are mentioned below |
| | TITLE | To search the keywords or developed search/query string only at the title of the article |
| | TITLE-ABS | To search the keywords or developed search/query string only at the title and abstract of the article |
| | TITLE-ABS-KEY | To search the keywords or developed search/query string only at the title, abstract and keywords of the article |
| | ALL | To search the keywords or developed search/query string only throughout the article |
| Phrase searching | "……" | To search for exact matches on words within a phrase search (with double quotes) |
| Truncation | Frequently used truncation symbols include the asterisk (*), a question mark (?) or a dollar sign ($). | Truncation is a searching technique used in databases in which placing a symbol at the end of the words. It aims to search for a word that could have multiple endings. |
| Wildcard | Frequently used wildcard symbols include the question mark (?), the pound/hatch symbol (#) or the asterisk (*). | Wildcard are symbols placed between words<br>Find variation of spelling (singular/plural)<br>Find variation of language spelling (American/British) |
| Boolean operator | OR | To include one or more of the terms (such as synonyms, related terms and variation) |
| | AND | To add terms and the terms may be far apart or to specify the search |
| | AND NOT | To exclude specific terms |

**Table 3** The sources and searching techniques Sources: Cooper et al. (2018)

| Step | The CRD Handbook | The Cochrane Handbook | Collaboration for environmental evidence | Joanna Briggs Institute reviewers manual | IQWiG Methods Resources | Systematic reviews in the social sciences: a practical guide | Eunethta | Campbell Handbook | Developing NICE guidelines: the Manual |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Searching electronic databases | Searching bibliographic databases | Searching online literature databases and catalogues | Databases (development of search strategies, phase one) | Bibliographic databases (1.search for primary literature. 2. search for SRs) | Databases | Bibliographic databases | Bibliographic databases (1. subject databases. 2. general databases) | No list of search methods but guidance distinguishes between database searching (first) and supplementary searching (second) |
| 2 | Scanning references lists of relevant studies | Handsearching | Searching websites of organisations and professional networks | Database searching (phase two) | Search in trial registries | Grey literature | Study registries | Conference proceedings and meeting abstracts | |
| 3 | Handsearching of key journals | Conference abstracts or proceedings | Searching the world wide web | Review references list | Clinical practice guideline databases and providers | identifying ongoing research | Searching for unpublished company documents | Existing review and publication reference lists | |
| 4 | Searching trials registers | Other reviews | Searching bibliographies of key articles/ reviews | Handsearching | Requests to manufacturers | Theses | Regulatory documents | Web searching | |

**Table 3** (continued)

| Step | The CRD Handbook | The Cochrane Handbook | Collaboration for environmental evidence | Joanna Briggs Institute reviewers manual | IQWiG Methods Resources | Systematic reviews in the social sciences: a practical guide | Eunethta | Campbell Handbook | Developing NICE guidelines: the Manual |
|---|---|---|---|---|---|---|---|---|---|
| 5 | Contacting experts and manufactures | We-searching | Contacting key individuals who work in the area | | Other data sources | Conference proceedings | Queries to authors | Unpublished studies | |
| 6 | Searching relevant internet resources | Unpublished and on-going studies (inc. researcher contact) | Citation searches for key papers/ included papers | | | Citation searching | Further search techniques | On-going studies | |
| 7 | Citation searching | | | | | Searching the web | | Institutional repositories | |
| 8 | Using a project website to canvas for studies | | | | | Contact with experts | | Handsearching | |
| 9 | | | | | | Trial registers | | | |

**Table 4** Explanation of manual searching techniques

| Techniques | Explanation |
|---|---|
| Handpicking | Is one of manual searching techniques that identify the articles/documents relevancy by searching journal content page by page and able to complete the non-indexing searching in the database |
| Backward tracking | Identification and examination of the references and works cited in an article/document |
| Forward tracking | Identification of articles that cite an original article/document after it had been published |
| Citation and reference tracking | A combination of forward and backward tracking. The researchers will first track the authors that had been cited in the text and then based on the citation, the researchers will track the full title of the articles written by the authors |

Database), PubMed, ScienceDirect, Scopus, TRID, Virtual Health Library, Web of Science (tested for Web of Science Core Collection, Medline), and Wiley Online Library. Cooper et al. (2018) suggested that researchers should not rely on database searching and diversify the sources and searching techniques (please refer to Table 3).

Google Scholars and Microsoft Academics are among the popular searching databases in the world (Athukorala et al. 2016; Gusenbauer 2019). However, these two databases are not suitable to be used as the principal database for SLR due to several problems and shortcomings (Gusenbauer and Haddaway 2020). Halevi et al. (2017) noted the problems related to the lack of quality control in Google Scholar, whereas Fagan (2017) and Housyar and Sotudeh (2018) are concerned on technical deficiencies such as less tolerance on complex search string and the lack of advanced search features. On the other hand, Haddaway et al. (2015) claimed that Google Scholar could act as a powerful supporting database in the searching process.

In addition to database searching, researchers should consider manual searching to maximise the list of literature. Some of the important practices in SLR are handpicking, backward tracking (or also known as reference searching), forward tracking (or also known as citation searching), citation and reference tracking (or also known as snowballing) to cross-check the available database (Shaffril et al. 2018) (please see Table 4). Xiao and Watson (2019) claimed that no database covers a complete set of published materials. Bates et al. (2017) claimed that several databases do not have perfect sensitivity on the keywords used by researchers which can be solved using manual searching in minimising the weaknesses. Thomas et al. (2017) suggest that the combination of manual and database searching can present extensive improvement of systematic reviews.

**3.3.1.3 When to stop searching?** There is no absolute answer on when to stop searching and determining whether the search is comprehensive and rigorous. For example, one might consider 30 articles as sufficient while some might consider it too little for SLR. Several scholars provided solutions such as Levy and Ellis (2006) on to stop when several searches using the same keywords in different databases produce no new result. Kastner et al. (2007) practised Capture-Mark-Recapture (CMR) technique to estimate the horizon of articles in the literature with their confidence interval.

The following are the key points to be considered in identification process

- Researchers enrich the main keywords in the identification process to retrieve more relevant articles for their SLR
- It is advantageous for researchers to master advanced search skills (Boolean operator, phrase searching, truncation, wildcard, field code function) in selected databases.
- Instead of relying on the database, researchers should rely on manual searching techniques (e.g., hand-searching and citation searching) to retrieve more articles.
- Gusenbauer and Haddaway (2020) suggested 14 databases that have the strength to act as the leading database in the article searching process.
- Google Scholars and Microsoft Academics can function as a supporting database for article searching and not the leading database due to several deficiencies.

### 3.3.2 Screening based on the inclusion and exclusion criteria

Screening is the second process in the systematic searching strategy that includes or excludes articles from the review and it is automatically assisted by the database. The process is based on the inclusion and exclusion criteria as determined by the researchers. The inclusion process uses the key features of the target population to answer research questions, whereas the exclusion process considers several characteristics of the population that might obstruct the study or increase the risks for undesired results that will be excluded from the researcher's consideration (Patino and Ferreira 2018).

**3.3.2.1 Selection of inclusion criteria** Several scholars tried to explain on inclusion and exclusion criteria for the SLR process. Xiao and Watson (2019) stated that it is better to have comprehensive criteria. The selected criteria must be able to categorise the research, provide reliable interpretation, and provide comprehensive works of literature. Okoli (2015) and Johnson and Hennessy (2019) stated that different situations would stimulate different needs, and there are no specific criteria in SLR. Kitchenham and Charters (2007) suggested that researchers should consider the criteria that can answer their research questions.

Linares-Espinos et al. (2018) mentioned that language, type of design, type of publication, and publication could overlap with each other. Okoli and Schabram (2010) focus on timeline publication as it is almost impossible to review all the published articles in the entire human history. Hence, researchers should determine the range of period for their review. To determine the best range of publication for their review, researchers might rely on the concept of study's maturity by Kraus et al. (2020). They stated that for a matured study where a good number of articles can tracked, the timeline publication might be shorter compared to a less matured study. Kraus et al. (2020) further explain that for a less matured study, a longer timeline publication is needed as there are limited number of articles and more scattered as a lot of research questions remain unanswered. Furthermore, Linares-Espinos et al. (2018) looked into publication language in which researchers should only select publications in languages that they understand because selecting articles from foreign languages can create confusion, add more costs for the review, and consume time. For the document type, Johnson and Hennessy (2019) suggested that researchers consider grey literature to obtain a comprehensive review. Grey literature is articles/documents that produced outside traditional publishing and distribution channels and it can be in a form of report, newsletter, reports, working papers, speeches, government documents, policy documents and others (Johnson and Hennessy 2019). There are critiques on the low quality of grey literature in which conclusion cannot be made until it is gathered and examined.

Delaney and Tamás (2018) explain the importance of grey literature in covering the existing literature for several fields of study.

In selecting appropriate criteria for the study, researchers might make common mistakes such as relying on similar variables when determining both inclusion and exclusion criteria. For example, a researcher might include studies from the Asian region and exclude countries from non-Asian regions. They might select criteria that are ineffective to answer the research questions and did not explain the key variables in the inclusion criteria when clarifying the external validity of the findings (Patino and Ferreira 2018). It is suggested to conduct a pilot study on all selected criteria to examine their suitability and maximise the potential of obtaining more relevant articles (Kitchenham and Charters 2007).

The following key points should be considered in the screening process

- Screening is the process that selects suitable and related articles for the review based on the inclusion and exclusion criteria as determined by the researchers.
- The criteria are diverse as different situations will stimulate different needs.
- Scholars reveal the importance of inclusion and exclusion criteria such as timeline publication, publication type, and language in their studies.

#### 3.3.3 Eligibility

Eligibility is a manual screening process. The identification and screening processes are assisted by computer and they are vulnerable to errors. Moher et al. (2009) and Liberati et al. (2009) state that researchers might include articles that do not conform to the criteria determined by the researchers after the screening process For example, there are possibilities to include studies which were conducted outside of Malaysia as the selected studies might be affiliated to researchers from Malaysia. The database can mistakenly identify them as studies conducted in Malaysia. For this issue, researchers can manually exclude them from consideration. The process focuses on reading the title, the abstracts, and if needed, the methodology section.

Key points to be considered in eligibility process

- Eligibility is an important manual process that enables researchers to minimise deficiencies made by the database.

### 3.4 Quality appraisal

#### 3.4.1 Tools that can assist researchers in appraisal of quality process

The remaining articles from the eligibility process need to be examined to ensure that the quality of methodology is free from any bias (Higgins et al. 2019). One of the most common ways to assess the quality of articles is to use tools, scales, checklist, or standard form. Siering et al. (2013) claimed that the most comprehensively validated assessment tool is the Appraisal of Guidelines for Research and Evaluation (AGREE II). They stated that it is necessary to rely on research questions when selecting the best assessment tool for the systematic review. Different researchers use different quality assessment tools in systematic reviews as they examine different quality criteria due to differences in research perspectives and needs for different systematic review settings (Seehra et al. 2016).

Higgins et al. (2019) created Cochrane risk-of-bias (RoB 2) tool to examine any bias, including biases that might occur in the randomisation process, changing the planned interventions, and missing outcome data that affect the results of randomised controlled trial. The Newcastle–Ottawa Scale is another example which examines the quality of articles based on eight items. It is further grouped into three categories as follows: 1) selecting the study groups; 2) determining the comparability of the groups; and 3) finding either the coverage or result for cohort or case–control studies. According to Hannes (2011), Critical Appraisal Skills Programme (CASP) is one of the most used quality assessments that has ten questions to detect methodological deficiencies. Pace et al. (2012) created Mixed Methods Appraisal Tool (MMAT) for the systematic review of mixed studies to ensure the value of diverse studies designs in a review. Hong et al. (2018) revised the Mixed Methods Appraisal Tool (MMAT) by including 25 criteria that cover five categories.

### 3.4.2 Approaches in assessing the quality of the articles

There are two approaches when assessing articles, namely qualitative and quantitative. The quality of articles can be examined quantitatively using an inter-rater agreement that refers to the score of the consistency given by the same person. There are other practices such as Cohen's Kappa analysis and concordance correlation coefficient (Petticrew and Roberts 2006). Researchers or experts can categorise the quality into three categories—low, moderate, and high. Reviews should select high-quality articles, followed by moderate quality. However, researchers should exclude low-quality articles during the consideration and only use them as foundational literature (Petticrew and Roberts 2006).

### 3.4.3 Selection of the reviewers

Besides that, it is important to select a suitable person to assess quality. It is suggested to hire at least two independent reviewers or the researchers can review the articles independently based on the criteria that they mutually agreed (Petticrew and Roberts 2006; Charrois 2015). They have to discuss and decide whether to maintain or exclude the articles that they failed to reach a mutual agreement regarding their quality (Petticrew and Roberts 2006). Besides that, researchers can communicate with corresponding researchers of the selected studies to gain clarification on the uncertainty in the methods and results (Charrois 2015).

Key points to be considered in the quality appraisal phase:

- Researchers should examine the quality of selected articles to ensure they are free from any bias and have a high level of quality.
- The systematic reviews use different quality assessment tools which provide different quality criteria depending on the need and nature of reviews.
- Each study design has its tools or scales or checklist to examine the quality of articles.
- Quality of articles can be assessed either qualitatively or quantitatively.
- At least two independent experts or two researchers can independently review the quality of articles based on the criteria.

### 3.5 Data extraction

The research question can guide data extraction as any data that can assist the researchers to answer the research questions can be extracted. Data extraction should be performed independently by two members to minimise errors in the data compilation process for analysis (Charrois 2015; Gomersall et al. 2015). Furthermore, a blinded review can minimise the bias when conducting the review. The team can discuss any disagreement on the extracted data in deciding whether to maintain or exclude them (Kitchenham and Charters 2007).

Key points to be considered in the data extraction phase:

- Researchers extract any data that assist them in answering their research questions.
- Data extraction should be performed independently by two researchers to minimise errors.

### 3.6 Data synthesis

Studies can be synthesised either quantitatively, qualitatively, or both. Quantitative synthesis measures can only be used on quantitative studies while qualitative synthesis can be applied to both quantitative and qualitative studies.

#### 3.6.1 Quantitative synthesis

Quantitative synthesis of review data is called meta-analysis which is a synthesis technique that focuses on the benefits and/or harms of treatment across multiple studies to offer the best estimate regarding the effect of an intervention (Morton et al. 2018). The researchers rely on established and systematic methods to determine the variances in sample size, variability in the results, and sensitivity of the findings on the protocol of systematic review (Shorten and Shorten 2013). Several issues need to be considered when conducting quantitative synthesis such as wide-range effects size, suspicion of publication, reporting bias, and effects of small sample size (Morton et al. 2018).

#### 3.6.2 Qualitative synthesis

Rousseau et al. (2008) equate the qualitative analysis to the creation via interpretation and explanation. Relativist epistemologies such as phenomenology or social construction are linked to interpretative and critical realist approaches. The data can be synthesis qualitatively using critical interpretive synthesis, ecological triangulation, fledging approaches, framework synthesis, grounded theory, meta-ethnography, meta-narrative review, meta-study, narrative summary, and thematic analysis. Table 5 presents the techniques in detail.

#### 3.6.3 Qualitative synthesis on mixed research designs

There is also a qualitative synthesis that combines both qualitative and quantitative studies. According to Soares et al. (2013) and Hopia et al. (2016), this is the best technique to solve an issue by looking from a diverse perspective. Prieto and Rumbo-Prieto (2018) consider it as the complete type of reviews to integrate methodologies and data for broader understanding.

**Table 5** Qualitative synthesis techniques

| Type of qualitative synthesis | Details |
|---|---|
| Narrative summary | Covers selection, chronicling, and ordering of evidence to result an account of the evidence (Green et al. 2006) |
| Thematic analysis | Identification of prominent or recurrent themes from the collected data of selected previous studies, and summarising these data under thematic headings (Braun and Clarke 2006) |
| Grounded theory | One of the technique to generate theory and it is 'grounded' in data that has been systematically composed and analysed (Noble and Mitchell 2016) |
| Meta-ethnography | An inductive, interpretive approach upon which most interpretive qualitative synthesis methods are based. This technique gear towards insights or clarifications that were not clear in the individual included studies (Paterson 2012) |
| Meta-study, | A combination of meta-data-analysis (focus of analysis in on the findings), meta-method (focus of analysis in on the methods) and meta-theory (focus of analysis is on the theory) (Patterson et al. 2001) |
| Meta-narrative review, | A meta-narrative review illuminates a heterogeneous topic area by highlighting the contrasting and complementary ways in which researchers have studied the same or a similar topic (Wong et al. 2013) |
| Critical interpretive synthesis | A combination of meta-ethnography and grounded theory that attempt to integrate multi-disciplinary and multi-method evidence (Dixon-wood et al. 2005) |
| Framework synthesis | The qualitative analysis will result in a large amount of data and causes difficulties for rigorous analysis, therefore, Brunton et al. (2006) have come out with framework synthesis idea which offers a highly structured approach for data management and analysis |
| Qualitative meta-summary | In qualitative meta-summary, the results are accumulated and summarised rather than 'transformed'. It is way of generating a 'map' of the contents of qualitative studies (Sandelowski et al. 2007) |

There are critiques on this kind of analysis such as varying epistemological approaches, and political and cultural contexts (Sandelowski et al. 2006; Mays et al. 2005) and the possibility of using narrative review, thematic analysis, meta-ethnography, and grounded theory in producing mixed research design (Barnett-Page and Thomas 2009). Flemming et al. (2018) concluded that thematic analysis is the most suitable in synthesising mixed research design and researchers need to be cautious on the use of framework synthesis. They stated that researchers need to be cautious for meta-ethnography when mixing research designs in a review. Dixon-wood et al. (2005) warn researchers on several disadvantages of qualitatively mixing research design such as the lack of transparency in narrative review, thematic analysis, and grounded theory, the lack of guidance for quality appraisal during studies inclusion in a review of grounded theory, and the inability of meta-study to explicitly cope with quantitative evidence.

Additionally, Whitemore and Knafl (2005) introduced integrative review that conducts a constant comparison in a broad array of qualitative designs to identify patterns based on the extracted data (quantitative and qualitative) and transform them into systematic categories, themes, variations, and relationships.

Key points to be considered in the data synthesis phase:

- Data can be analysed either quantitatively or qualitatively.
- Quantitative synthesis of the review data is also known as meta-analysis.
- Some of the qualitative synthesis techniques in SLR are grounded theory, metaethnography, narrative summary, and thematic analysis.
- Thematic analysis is the most suitable in synthesising mixed research design while researchers need to be cautious for framework synthesis.

### 3.7 Data demonstration

#### 3.7.1 Referring to a publication or reporting standards to demonstrate the review

PRISMA and RAMESES can suggest things that should be demonstrated in SLR articles besides being a guide for systematic review. There is the need to explain on systematic literature review to provide adequate information for others. The details will allow future researchers to replicate the entire procedure in their studies.

#### 3.7.2 Use of advanced search string, inclusion of flow diagrams and developing table of findings

Furthermore, researchers should try to ensure that the SLR process is clear and the data is supported by conclusions. The search string should utilise all possible coding symbols to ensure rigorous searching in retrieving more articles, and all inclusion and exclusion criteria should be described and justified (Peters et al. 2015). The flow diagram can be used to describe the A–Z process of SLR in an organised manner. Among the established flow diagrams are PRISMA flow diagram (Moher et al. 2009) and customised flow diagram (Shaffril et al. 2019). Tables related to search string, inclusion and exclusion criteria can be included. It is suggested that the table of findings should be one of the main priorities in SLR. The table offers a simple explanation and understanding of the results. One of the aims of the literature review is to identify the patterns of previous studies which can be done using SLR. SLR can identify research gaps and offer direction for future studies (Petticrew and Roberts 2006). Researchers can proceed with their literature review after confirming all of these criteria.

Key points to be considered in the data demonstration phase:

- PRISMA and RAMESES can suggest things to be demonstrated in the SLR articles.
- Each SLR section should provide detailed information.
- A flow diagram can be included in SLR.
- Table of findings offers a simple explanation and understanding of the SLR results.

## 4 Conclusion

SLR have been well advanced in the field of medical or health related. Understandably, in recent years, there are increasing number of researchers, especially those from non-health field of studies who placed their interests on this kind of review. However, lacking of methodological references in SLR and less suitability of existing methodological guidance are making SLR of non-health related fail to fully abide to standard guidance which lead to

limited penetration and use of systematic approaches. The main objective of this paper is to provide a general understanding of basic methodologies for SLR along with several options or alternatives that can be considered by the non-health related scholars in their efforts to develop their systematic literature review. To develop this methodological-based papers, it was based on information provided by 75 selected articles. The articles were selected from several leading sources such Scopus, Science Direct, Google Scholars, and Google engine search. All of the selected articles were appraised its quality by two reviewers. The guidance is based on seven main aspects of SLR methodology as follows: (1) the development and validation of the review protocol/publication standard/reporting standard/guidelines; (2) the formulation of research questions; (3) systematic searching strategies; (4) quality appraisal; (5) data extraction; (6) data synthesis; and (7) data demonstration.

The methodological process in SLR should offer a complete guide for future scholars. SLR starts with developing and validating the review protocol/publication standard and reporting standard/guidance which are manuals of systematic plans that guide researchers on things that should be considered in the review. The next step is formulating research questions to guide the entire SLR process. Both specific and broad research questions have their advantages.

The next process is systematic searching strategies which have three sub-processes: identification, screening, and eligibility. Identification is the process to enrich the main keywords. Instead of using the main keywords, researchers need to identify their synonyms, related terms, and variation. More keywords basically will produce more related potential articles for the review. Researchers need to select the most appropriate and related database instead of relying on database searching by diversifying their sources and searching techniques. It is believed that researchers who are equipped with search string/query development skills can perform more rigorous searching. The second step is the screening process that covers the inclusion and exclusion criteria in the review. Different situations have different needs, and the selected criteria must match the research questions.

The last process is the eligibility, which is a manual screening process to minimise the deficiencies in the database. Before data extraction and analysis, the quality of selected articles needs to be assessed either qualitatively or quantitatively using tools, scales, checklist, or standard form. The remaining articles can be extracted for relevant data before the analysis process. The data can be analysed either using fully quantitative (meta-analysis), fully qualitative, or a mix of qualitative and quantitative approaches.

PRISMA and RAMESES can suggest standard things to be considered in the publication. Any SLR should offer in-depth details to offer guidance for future scholars to replicate the entire procedure in their studies. In addition, any SLR should consider the inclusion of flow diagram, table of search string, table of inclusion and exclusion criteria.

## References

Athukorala, K., Głowacka, D., Jacucci, G., Oulasvirta, A., Vreeken, J.: Is exploratory search different?: A comparison of information search behavior for exploratory and lookup tasks. J. Assoc. Inf. Sci. Technol. **67**(11), 2635–2651 (2016). https://doi.org/10.1002/asi.23617

Barnett-Page, E., Thomas, J.: Methods for the synthesis of qualitative research: a critical review. BMC Med. Res. Methodol. (2009). https://doi.org/10.1186/1471-2288-9-59

Bates, J., Best, P., McQuilkin, J., Taylor, B.: Will web search engines replace bibliographic databases in the systematic identification of research? J. Acad. Librariansh. **43**(1), 8–17 (2017)

Berrang-Ford, L., Pearce, T., Ford, J.D.: Systematic review approaches for climate change adaptation research. Reg. Environ. Change **15**(5), 755–769 (2015). https://doi.org/10.1007/s10113-014-0708-7
Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
Burgers, C., Brugman, B.C., Boeynaems, A.: Systematic literature reviews: four applications for interdisciplinary research. J. Pragmat. **145**, 102–109 (2019)
Brunton, G., Oliver, S., Oliver, K., Lorenc, T.: A Synthesis of Research Addressing Children's, Young People's and Parents' Views of Walking and Cycling for Transport. EPPI-Centre, Social. Science Research Unit, Institute of Education, University of London, London (2006)
Cañón, M., Buitrago-Gómez, Q.: The research question in clinical practice: a guideline for its formulation. Rev Colomb Psiquiatr. **47**(3), 193–200 (2018). https://doi.org/10.1016/j.rcp.2016.06.004
Centre for Reviews and Dissemination.: Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. University of York, York (2006)
Charrois, T.L.: Systematic reviews: What do you get to know to get started? Can. J. Hosp. Pharm. **68**(2), 144–148 (2015)
Cooke, A., Smith, D., Booth, A.: Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual. Health Res. **22**(10), 1435–1443 (2012)
Cooper, C., Booth, A., Campbell, J., Britten, N., Garside, R.: Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Med. Res. Methodol. **18**, 85 (2018)
Creswell, J.: Qualitative Inquiry & Research Design: Choosing Among Five Approaches, 3rd edn. Sage Publications, Inc., Thousand Oaks, CA (2013)
del Amo, I.F., Erkoyuncu, J.A., Roy, R., Palmarini, R., Onoufriou, D.: A systematic review of augmented reality content-related techniques for knowledge transfer in maintenance applications. Comput. Ind. **103**, 47–71 (2018)
Delaney, A., Tamás, P.A.: Searching for evidence or approval? A commentary on database search in systematic reviews and alternative information retrieval methodologies. Res. Synth. Method **9**(1), 124–131 (2018)
Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., Sutton, A.: Synthesising qualitative and quantitative evidence: a review of possible methods. J. Health Serv. Res. Policy **10**(1), 45–53 (2005)
Doody, O., Bailey, M.E.: Setting a research question, aim and objective. Nurse Res. **23**(4), 19–23 (2016)
Durach, C.F., Kembro, J., Wieland, A.: A new paradigm for systematic literature reviews in supply chain management. J. Supply Chain Manag. **53**(4), 67–85 (2017)
Fagan, J.C.: An evidence-based review of academic web search engines, 2014–2016: implications for Librarians' Practice and Research Agenda. Inf. Technol. Libr. **36**(2), 7–47 (2017)
Flemming, K., Booth, A., Garside, R., Tunc¸alp, O., Noyes J.: Qualitative evidence synthesis for complex interventions and guideline development: clarification of the purpose, designs and relevant methods. BMJ Global Health (2018)
Gomersall, J.S., Jadotte, Y.T., Xue, Y., Lockwood, S., Riddle, D., Preda, A.: Conducting systematic reviews of economic evaluations. Int. J. Evid.-Based Healthc. **13**(3), 170–178 (2015). https://doi.org/10.1097/XEB.0000000000000063
Green, B.N., Johnson, C.D., Adams, A.: Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. J. Chiropr. Med. **5**(3), 101–117 (2006). https://doi.org/10.1016/S0899-3467(07)60142-6
Greyson, D., Rafferty, E., Slater, L., MacDonald, N., Bettinger, J.A., Dubé, È., MacDonald, S.E.: Systematic review searches must be systematic, comprehensive, and transparent: a critique of Perman et al. BMC Public Health **19**(1), 1–6 (2019). https://doi.org/10.1186/s12889-018-6275-y
Gusenbauer, M.: Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Sciencetometrics **118**(1), 177–214 (2019)
Gusenbauer M, Haddaway NR (2020) Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res. Synth. Methods. 11(2):181–217. https://doi.org/10.1002/jrsm.1378
Haddaway, N.R., Collins, A.M., Coughlin, D., Kirk, S.: The role of google scholar in evidence reviews and its applicability to grey literature searching. PLoS ONE **10**(9), e0138237 (2015). https://doi.org/10.1371/journal.pone.0138237
Haddaway, N.R., Macura, B., Whaley, P., Pulin, A.S.: ROSES Reporting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environ. Evid **7**, 7 (2018). https://doi.org/10.1186/s13750-018-0121-7

Halevi, G., Moed, H., Bar-Illan, J.: Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation. Revi Lit **11**(3), 823–834 (2017)
Hannes, K.: Critical appraisal of qualitative research. In: Noyes, J., Hannes, K., Harden, A., Harris, J., Lewin, S., Lockwood, C. (eds.) Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions. Cochrane Collaboration Qualitative Methods Group, London (2011)
Higgins, J.P.T., Altman, D.G., Gotzsche, P.C., Juni, P., Moher, D., Oxman, A.D., Savovic, J., Schulz, K.F., Weeks, L., Sterne, J.A.C.: The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ **343**(7829), 1–9 (2011). https://doi.org/10.1136/bmj.d5928
Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (eds.): Cochrane Handbook for Systematic Reviews of Interventions, 2nd edn. Wiley, Chichester (UK) (2019)
Hong, Q.N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M-C., Vedel, I.: Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada (2018)
Housyar, M., Sotudeh, H.: A reflection on the applicability of Google Scholar as a tool for comprehensive retrieval in bibliometric research and systematic reviews. Int. J. Inf. Sci. Manag. **16**(2), 1–17 (2018)
Hopia, H., Latvala, E., Liimatainen, L.: Reviewing the methodology of an integrative review. Scand. J. Caring Sci. **30**(4), 662–669 (2016)
Johnson, B.T., Hennessy, E.A.: Systematic reviews and meta-analyses in the health sciences: best practice methods for research syntheses. Soc. Sci. Med. **233**, 237–251 (2019)
Kastner, M., Straus, S., Goldsmith, C.H.: Estimating the horizon of articles to decide when to stop searching in systematic reviews: an example using a systematic review of RCTs evaluating osteoporosis clinical decision support tools. AMIA Annu. Symp. Proc. Arch. **2007**, 389–393 (2007)
Kitchenham, B.A., Charters, S.M.: Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report (2007)
Kraus, S., Breier, M., Dasí-Rodríguez, S.: The art of crafting a systematic literature review in entrepreneurship research. Int. Entrep. Manag. J. **16**(3), 1023–1042 (2020). https://doi.org/10.1007/s11365-020-00635-4
Kushwah, S., Dhir, A., Sagar, M., Gupta, B.: Determinants of organic food consumption. A systematic literature review on motives and barriers. Appetite **143**, 104402 (2019)
Levy, Y., Ellis, T.J.: A systems approach to conduct an effective literature review in supports of information system research. Inf. Sci. J. **9**, 181–212 (2006)
Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med **6**(7), e1000100 (2009)
Linares-Espinós, E., Hernández, V., Domínguez-Escrig, J.L., Fernández-Pello, S., Hevia, V., Mayor, J., Padilla-Fernández, B., Ribal, M.J.: Methodology of systematic review. Actas urologicas españolas **42**(8), 499–506 (2018)
Lockwood, C., Munn, Z., Porritt, K.: Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int. J. Evid. Based Healthc. 13(3), 179–187 (2015). https://doi.org/10.1097/XEB.0000000000000062
Long, A.F., Godfrey, M.: An evaluation tool to assess the quality of qualitative research studies. Int. J. Soc. Res. Methodol. **7**(2), 181–196 (2004)
Mallet, R., Hagen-Zanker, J., Slater, R., Duvendack, M.: The benefits and challenges of using systematic reviews in international development research. J. Dev. Eff. **4**, 445–455 (2012)
Mantzoukas, S.: Facilitating research students in formulating qualitative research questions. Nurse Educ. Today **28**(3), 371–377 (2008)
Mays, N., Pope, C., Popay, J.: Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. J. Health Serv. Res. Policy **10**(1), 6–20 (2005)
Mengist, W., Soromessa, T., Legese, G.: Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX **7**, 100777 (2020). https://doi.org/10.1016/j.mex.2019.100777
Methley, A.M., Campbell, S., Chew-Graham, C., McNally, R., Cheraghi-Sohi, S.: PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv. Res. (2014). https://doi.org/10.1186/s12913-014-0579-0
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., PRISMA Group: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med **21;6**(7), e1000097 (2009)

Morton SC, Murad MH, O'Connor E, Lee CS, Booth M, Vandermeer BW, Snowden JM, D'Anci KE, Fu R, Gartlehner G, Wang Z, Steele DW (2018) Quantitative synthesis—an update. methods guide for comparative effectiveness reviews. (Prepared by the Scientific Resource Center under Contract No. 290-2012-0004-C). AHRQ Publication No. 18-EHC007- EF. Rockville, MD: Agency for Healthcare Research and Quality. Posted final reports are located on the Effective Health Care Program search page. https://doi.org/10.23970/AHRQEPCMETHGUIDE3

Noble, H., Mitchell, M.: What is grounded theory? Evid. Based Nurs. **19**(2), 34–35 (2016). https://doi.org/10.1136/eb-2016-102306

Okoli, C.: A guide to conducting a standalone systematic literature review. Commun. Assoc. Inf. Syst. **37**, 879–910 (2015)

Okoli, C., Schabram, K.: A guide to conducting a systematic literature review of information systems research. Philos. Methodol. Econ. eJ. (2010). https://doi.org/10.2139/ssrn.1954824

Onwuegbuzie, A.J., Leech, N.L.: Sampling designs in qualitative research: making the sampling process more public. Qual. Rep. **12**(2), 238–254 (2007)

Pace, R., Pluye, P., Bartlett, G., Macaulay, A.C., Salsberg, J., Jagosh, J., Seller, R.: Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int. J. Nurs Stud. **49**(1), 47–53 (2012). https://doi.org/10.1016/j.ijnurstu.2011.07.002

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T., Mulrow, C.D., Shamseer, L., Moher, D.: Mapping of reporting guidance for systematic reviews and meta-analyses generated a comprehensive item bank for future reporting guidelines. J. Clin. Epidemiol. **118**, 60–68 (2020)

Paterson, B.L., Thorne, S.E., Canam, C., Jillings, C.: Meta-Study of Qualitative Health Research. A Practical Guide to Meta-Analysis and Meta-Synthesis. Sage Publications, Thousand Oaks (2001)

Palaskar, J.N.: Framing the research question using PICO strategy. J. Dent. Allied Sci. **6**(2), 55 (2017)

Patino, C.M., Ferreira, J.C.: Inclusion and exclusion criteria in research studies: definitions and why they matter. J. Bras. Pneumol. **44**(2), 84 (2018)

Peters, M.D., Godfrey, C.M., Khalil, H., McInerney, P., Parker, D., Soares, C.B.: Guidance for conducting systematic scoping reviews. Int. J. Evid.-Based Healthc. **13**(3), 141–146 (2015). https://doi.org/10.1097/XEB.0000000000000050

Petticrew, M., Roberts, H.: Systematic Reviews in the Social Sciences: A Practical Guide. Blackwell Publishing Ltd, Oxford (2006)

Prieto and Rumbo-Prieto: The systematic review: plurality of approaches and methodologies. Enferm. Clín. (English Edition) **28**(6), 387–393 (2018)

Reim, W., Parida, V., Örtqvist, D.: Product-Service Systems (PSS) business models and tactics – a systematic literature review. J. Clean. Prod. **97**, 61–75 (2015). https://doi.org/10.1016/j.jclepro.2014.07.003

Robinson, P., Lowe, J.: Literature reviews vs systematic reviews. Aust. N. Z. J. Public Health **39**(2), 103 (2015)

Rousseau, D. M., Manning, J., Denyer, D.: Evidence in management and organizational science: assembling the field's full weight of scientific knowledge through syntheses. In: AIM Research Working Paper Series: Advanced Institute of Management Research (2008)

Sandelowski, M., Barroso, J., Voils, C.I.: Using qualitative metasummary to synthesize qualitative and quantitative descriptive findings. Res. Nuurs. Health **30**(1), 99–111 (2007). https://doi.org/10.1002/nur.20176

Sandelowski, M., Voils, C.I., Barroso, J.: Defining and designing mixed research synthesis studies. Res. Schools: Nat. Ref. J. Spons. Mid-South Educ. Res. Assoc. Univ. Alabama **13**(1), 29 (2006)

Schardt, C., Adams, M.B., Owens, T., Keitz, S., Fontelo, P.: Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med. Inf. Decis. Mak. **7**, 16 (2007). https://doi.org/10.1186/1472-6947-7-16

Seehra, J., Pandis, N., Koletsi, D.: Use of quality assessment tools in systematic reviews was varied and inconsistent. J. Clin. Epidemiol. **69**, 179–184 (2016)

Shaffril, H.A.M., Krauss, S.E., Samsuddin, S.F.: A systematic review on Asian's farmers' adaptation practices towards climate change. Sci. Total Environ. **644**, 683–695 (2018)

Shaffril, H.A.M., Abu Samah, A., Samsuddin, S.F., Ali, Z.: Mirror-mirror on the wall, what climate change adaptation strategies are practiced by the Asian's fishermen of all? J. Clean. Prod. **232**, 104–117 (2019)

Shorten, A., Shorten, B.: What is meta-analysis. Evid. Based Nurs. **16**(1), 3–4 (2013)

Siering, U., Eikermann, M., Hausner, E., Hoffmann-Eßer, W., Neugebauer, E.A.: Appraisal tools for clinical practice guidelines: a systematic review. PLoS ONE **8**(12), e82915 (2013)

Soares, C.B., Hoga, L., Sangaleti, C., Yonekura, T., Peduzzi, M., Silva, D.: Integrative review in nursing research and EBP: a type of systematic review? Int. J.Evid.-Based Healthcare **11**(3), 246–247 (2013)

Thomas, J., Noel-Storrb, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilta, I., Synnote, A., Turnere, T., Elliott, J.: Living systematic reviews: 2. Combining human and machine effort. J. Clin. Epidemiol. **91**, 31–37 (2017)

Thomas, J., Kneale, D., McKenzie, J.E., Brennan, S.E., Bhaumik, S.: Chapter 2: Determining the scope of the review and the questions it will address. In: Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (eds.) Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (2020). Cochrane (2020). https://www.training.cochrane.org/handbook

Wanden-Berghe, C., Sanz-Valero, J.: Systematic reviews in nutrition: standardized methodology. Br. J. Nutr. **107**, S3–S7 (2012)

Whittemore, R., Knafl, K.: The integrative review: updated methodology. J. Adv. Nurs. **52**(5), 546–553 (2005)

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., Pawson, R.: RAMESES publication standards: realist syntheses. BMC Med. **11**, 21 (2013). https://doi.org/10.1186/1741-7015-11-21

Xiao, Y., Watson, M.: Guidance on conducting a systematic literature review. J. Plan. Educ. **39**(1), 93–112 (2019)

Younger, P.: Using Google Scholar to conduct a literature search. Nurs. Stand. **24**(45), 40–46 (2010)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.