**Choosing the right tool for the job: Screening tools for systematic reviews in education**

*Qiyang Zhang*

*and*

*Amanda Neitzel*

School of Education, Johns Hopkins University

**Correspondence**

Qiyang Zhang

2800 N Charles Street, Baltimore, MD21218, USA

**How to cite this preprint:** Zhang Q., & Neitzel A., (2023). Choosing the right tool for the job: Screening tools for systematic reviews in education. arXiv.

**Abstract**

In recent years, the rapid development of artificial intelligence has enabled the launch of many new screening tools. This review aims to facilitate screening tool selection through a systematic narrative review and feature analysis. The current adoption rate of transparent tool reporting is low: by screening 191 studies published in the *Review of Educational Research* since 2015, we found that only eight studies reported screening tools. More research is needed to understand reasons behind this phenomenon. After consulting various sources, 26 available screening tools in the market were found. Among them, we identified and evaluated 12 screening tools for educational reviewers and ranked them in descending order of feature score: Covidence (1), DistillerSR (2, tied), EPPI-Reviewer (2, tied), CADIMA (4), Swift-Active (5), Rayyan (6, tied), SysRev (6, tied), Abstrackr (8, tied), ReLiS (8, tied), RevMan (8, tied), ASReview (11), and Excel (12). In discussion, we provide insights on the promise and bias in tools’ machine learning algorithms. Our results encourage researchers to report their tool usage in publications and select tools based on suitability instead of convenience.

**KEYWORDS: systematic reviews, screening tools, machine learning, feature analysis, decision tree**

**Choosing the right tool for the job: Screening tools for systematic reviews in education**

*If I had eight hours to chop a tree, I’d spend six sharpening my ax.*

*-Abraham Lincoln*

*History of Tool Development in Systematic Review*

In systematic reviews, the *screening process* is the procedure for identifying eligible studies for inclusion and analysis. A screening process may include two stages, starting with a title and abstract screening stage where the studies are assessed for relevance and a second stage where the full-text of the article is reviewed for inclusion. The standard practice of screening involves multiple reviewers making inclusion-exclusion decisions on the same set of studies as a consistency check (Cooper et al., 2019). Historically, systematic reviewers often manually collaborated to screen studies using hard copies. A folder of printed-out potential studies passed around an office room often underpinned the basis for early systematic reviews and meta-analyses. This manual screening process took a considerable amount of time to manage literature and precluded blind review as well as multi-geographical collaboration. Fortunately, technological development has transformed screening routines, drastically saving time and labor committed to this intensive process.

With Excel’s launch in 1985, reviewers gradually started to digitize the screening and coding process using spreadsheets. In the early 2000s, citation management software, such as EndNote and Zotero, started to provide a more convenient platform for literature storage and retrieval. These reference management tools began to replace spreadsheets. There has also been the development of tools specifically designed to support the screening and reviewing stages of systematic reviews. In 2010s, the rapid boom in artificial intelligence (AI) and machine-learning applications has prompted the launch of some new meta-analysis software tools with AI-enhanced features.

*Why Screening Tools are better than Spreadsheets and Citation Management Tools*

Despite the convenience and benefits spreadsheets and citation management tools brought, they have major limitations in essential features because they are not developed specifically for screening purposes. Excel spreadsheets were primarily developed for task management, data organization, analysis, and visualization purposes (Microsoft, 2021) and citation management tools were developed to collect, organize, annotate, cite, and share research. Systematic reviewers and meta-analysts who rely on spreadsheets and citation management tools face various problems in conducting recommended practices. For instance, best practices in screening abstracts include ensuring independent double-screening, conducting pilot testing of screening, and arranging regular team meetings to reconcile disagreements (Polanin et al., 2019). Screening tools developed based on these guidelines ideally embed features such as blind reviewing, screening progress overview, conflict resolution, inter-rater reliability, and reviewers’ progress tracking. Domain knowledge enabled the screening tool developers to incorporate protocols for high-quality systematic reviews. These added functions partially explain the increasing popularity of screening tools among systematic review teams.

The other reason for review teams to switch to screening tools is that advanced algorithms can help expedite the often manual and tiresome screening process. Empowering systematic review teams to screen faster is the main attraction of screening tools. For example, some new screening tools were equipped with machine learning-based text classification algorithms to rank the relevancy of the studies based on human reviewers’ screening patterns, which significantly reduce resources required for conducting reviews (Cohen et al., 2006). These features are particularly helpful when conducting large-scale reviews in cross-disciplinary fields such as education. The additional functions enabled by these semi-automated software tools hold promise to save a tremendous amount of time and labor in evidence-based synthesis (Ouzzani et al., 2016; Rathbone et al., 2015), if selected and used appropriately.

The availability of new screening tools present new problems. Not all screening tools are the same and researchers’ selection of the tools can be consequential. For instance, some tools may have less accurate deduplication functions (McKeown & Mir, 2021) while some other tools may have biased algorithms embedded in machine learning features (Varghese et al., 2019). As readers, information on tool usage behind systematic reviews enable us to better compare different reviews of the same topic or interpret the review results. As researchers, understanding reasons behind tools used in past reviews can guide us select suitable tools and conduct better research. In published reviews, the screening process is generally not well-explained and the rationale behind tool selection is rarely transparent. This article aims to provide information for educational researchers to select the right tool for a particular set of needs. We will identify and compare software tools specifically designed to perform title and abstract screening and/or full-text review. Meanwhile, we will assess published review papers in education to understand the current tool usage among educational researchers. Eventually, our goal is to provide sufficient and simple information, such as a decision tree, to assist reviewers’ tool selection.

*Screening Tools in Education*

Within the field of education, there is a lack of research comparing different screening tools. This kind of research has been conducted in some other fields, such as biomedical research (Van der Mierden et al., 2019) and healthcare (Harrison et al., 2020). Van der Mierden et al. (2019) performed feature analysis on 16 available screening tools in biomedical research and ranked Rayyan as the best free tool and Microsoft products as the least preferable tools for biomedical researchers. Harrison et al. (2020) conducted a weighted feature analysis on six screening tools in healthcare and recommended Covidence and Rayyan for healthcare systematic reviewers. However, considering this question from the perspective of education is valuable. While the screening process (title and abstract screening followed by full-text review against inclusion/exclusion criteria) is universal and some features such as the ability to collaborate by allowing multiple users would be of interest regardless of field, there are specific features discussed below that would be particularly helpful within education. We build on similar work in other fields (i.e., Van der Mierden et al., 2019) by including an additional focus on these features of particular relevance in education.

Evidence-based reform in education has contributed to a proliferation of well-designed and large-scale experiments in recent years, with a corresponding call for up-to-date, comprehensive systematic reviews across a broad set of topics. However, this proliferation can be regarded as a double-edged sword. On one hand, more evidence contributes to the development of rigorous educational research synthesis. On the other hand, a higher volume of studies can make meta-analysis more labor- and time-consuming. In addition, education is an interdisciplinary field that requires literature search from multiple databases and sources. Educational systematic reviewers urgently need technology-based screening tools with time-saving features, such as deduplication and machine learning, to expedite the screening process. These needs may differ in other fields, where they need to only search a single or several databases and the results will only include a few hundred studies. In the field of education, there is a particularly acute urgency for tools that can make screening less resource-intensive since conducting interdisciplinary searches often retrieves thousands of studies to screen. In large educational systematic reviews, machine learning approaches are particularly helpful. Since machine learning is of special interest, we propose to take a deeper look at the approaches used in machine learning, so that it is not just whether or not machine learning is available, but whether it also includes more modern deep neural networks.

The speed with which the education field develops, tests, and markets new approaches is fast. Vendors and developers are constantly bringing to market new programs and practices that school districts must decide whether to adopt. The results of these must be rapidly integrated into existing reviews, so that the ability to update existing reviews is an important feature in education that may not be relevant in other fields, where the pace of testing and change is slower or more regulated by other forces.

Another factor that may determine which features are especially relevant is domain-specific definitions of what constitutes a “high-quality” review. These differences in what is “good enough” are expected to vary if we accept a best-evidence approach, which focuses the synthesis on the best available evidence (Slavin, 1986). It is natural to believe that reviews in fields with more randomized control trials would focus on those for their reviews, while fields with fewer randomized studies or fields for which randomized studies are not always appropriate or possible would include a more diverse set of research designs. Within education, the level of evidence available depends on the particular question of interest, so there must be flexibility in how or even whether those standards are handled during the screening process or instead are addressed during the coding process. The process of assessing the research design could be included in the software tool if reviewers wish to focus only on including studies that meet a particular standard of methodological rigor. The tools that are designed for fields that prize experiments may not be able to flexibly handle additional study designs, such as quasi-experiments or case studies.

Thirdly, there are even particular software tools that were developed for specific fields that may struggle to be adapted to the education field. For example, Parsifal, a screening tool developed for the software engineering field, is fully integrated with one database, Scopus, so that one could conduct searches of that database from within the software, save the results of the search, and then import to the screening tool. This is a remarkable feature that simplifies the process but has not yet been implemented with education-specific databases such as ERIC. Therefore, while integration of databases directly into the screening software is of value in some fields, it is less relevant in education when those integrated databases are not those with the most coverage of education-related topics. This ability is even of less relevance for educational researchers because education is an interdisciplinary field, requiring the use of many different databases, making the integration of a single database less useful. Because of these differences, the conclusions drawn from biomedical research and healthcare on choosing screening tools are not necessarily transferable to educational researchers. Therefore, we are motivated to conduct this study specifically in the field of education.

This study intends to identify screening tools used by educational researchers and provide guidance for tool selection in the field of education. To the best of our knowledge, this study is the first to address this issue for educational researchers. Through systematic narrative review and feature analysis of tools’ features, this paper aims to examine currently available screening tools in education to empower educational meta-analysts with the information necessary to optimize tool selection. To facilitate the selection process of screening tools, we present an opinion-based decision tree at the end of the paper.

*Review Objectives*

The three objectives of this study are to:

1. Identify available screening tools for educational analysts. To do so, we will first 1a) identify all available screening tools in the market, and then 1b) identify the subset of those screening tools that could be used by educational researchers.
2. Assess the tools educational researchers are already using through tool reporting in published reviews. To achieve this objective, we used two steps: 2a) screening published articles in *Review of Educational Research (RER)* to determine the prevalence of reporting screening tools in published educational research, and 2b) cross-checking whether we have identified a complete set of screening tools used by educational researchers in Objective 1b.
3. Analyze differing features of available screening tools to inform educational meta-analysts’ tool selection. In this objective, we have two procedures: 3a) collecting information on screening tools identified in Objective 1b, and 3b) ranking tools through feature analysis to inform educational systematic reviewers’ tool selection.

**Method**

*Objective 1. Identifying available screening tools for educational analysts*

In the current screening tool software market, there are many meta-analytical tools developed for specific purposes in particular fields. For example, SyRF was developed for preclinical studies; SRDB.PRO was developed for the pharmaceutical industry and healthcare consultancies; Colandr was launched by biodiversity researchers; RobotAnalyst and SRDR Plus were developed for healthcare researchers; and PASRIFAL was developed for software engineering. As discussed in the previous section, not all screening tools are feasible for educational systematic research due to niche differences across fields. To locate screening tools currently available in the market, we conducted a web-based search to obtain names of different tools from academic publications (Harrison et al., 2020; Schoot et al., 2021; Van der Mierden et al., 2019), personal blogs (Bradburn, 2018), university library resources (Roth, 2021), and research centers (Center for Evidence Synthesis in Health, 2021). In order to narrow down and identify available and appropriate tools for educational researchers, we included and analyzed the screening tools that have been used in educational research. To assess this criterion, we searched for “tool name + education/school/students + meta-analysis/systematic review” on Google Scholar to find systematic reviews in education that cite the screening tools. Finally, among all the identified tools, we exclude tools that were designed for a specific non-education field and tools that are coding frameworks or R packages.

*Objective 2. Assessing the tools educational researchers are already using through tool reporting in published reviews*

Variations in reviewers’ screening procedures and usage of different screening tools can result in systematic differences in their research results. With the advancement of the open science movement and transparent research, more researchers are expected to report their screening procedure in detail (Patall, 2021). We are interested in finding out the current status of reporting the tools used for screening in educational research. Therefore, we screened publications from the journal *— Review of Educational Research (RER) —* to conduct a systematic search and screening. *RER* is among the most impactful and representative peer-reviewed journals for systematic reviews in the field of education. Because almost all new screening software tools, except Revman and Excel, were released post-2010 (see Table 1) and it takes time for research teams to train with and implement new software effectively, we selected the cut-off year of 2015 to assess published studies’ reporting status of screening tools. This also ensures that we are examining the tools currently in-use in the field and most likely to still be available.

Article retrieval was enabled with an open-source web-app called Paperfetcher (Pallath & Zhang, 2022), which automates handsearching, forward and backward citation chasing, and allows bulk export to common bibliographic data formats used by screening tools. We used the handsearching function in Paperfetcher to retrieve studies published in *RER* from January 1st, 2015, to October 2nd, 2021. The screening process involves two steps. First, we examine each study’s eligibility by excluding studies that do not report any tool usage in the method section. Second, in the full-text screening, we classified studies by their reported use of tools used during the screening process. In the second step, we only included studies that report use of screening tools. We chose Covidence to conduct this screening process because it enables both title and abstract screening and full-text review, and because we have an institutional license to use this tool.

*Objective 3. Analyzing differing features of available screening tools to inform educational meta-analysts’ tool selection*

To understand specific features and functions of our identified screening tools, we conducted a feature analysis. The process of selecting and assigning scores to features was guided by the Kitchenham’s (1996) DESMET method, a validated method to evaluate a software (Kitchenham & Linkman, 2000). Out of all the evaluation methods presented in DESMET, we selected screening mode under feature analysis since this method requires the least amount of labor and is the fastest and the least costly method. The first step in the screening mode approach is to identify candidate tools for evaluation, which is completed in Objective 1b. The second step is to choose features, which is completed after an iterative process of discussion among the authors and within the authors’ internal network. In addition, previous research in similar topics and the PRISMA statement are consulted and considered in the feature selection process. After deciding on which features to compare, we then examined these screening tools’ official websites, publicly-accessible training videos, and the academic publications introducing the tools. Tools that were developed by academics gave detailed documentation on their functions, robustness, and limitations in publication. There was no restriction on the date of development of tools – tools could have been developed at any point.

Tools’ features were coded as dummy variables (0/1), where 0 means absence of a feature and 1 means presence of a feature. We separated various features into general functions and the two stages of the screening process: title and abstract screening stage, and full-text review stage. General functions include bulk import and export, teamwork, blind review, duplicate removal, and research update. Title and abstract screening stage’s features include: possibility of title/abstract screening, machine learning classifiers, deep learning, and inter-rater reliability. Apart from machine learning classifiers, we also coded whether the screening tool uses deep neural network models for improved performance. Full-text review stage’s features include: possibility of full-text review, inter-rater reliability, and reason labels. Categorizing these features gives reviewers a more holistic comparison for each stage of the systematic reviewing process and enables them to compare and select the most suitable tool. We recorded some other non-quantifiable features separately, such as cost and privacy policy. We also calculated a feature analysis score that is simply the sum of the number of features coded above. While this approach treats each feature as equally important, we recognize this is likely not the case, because each researcher will rank them differently. However, this does give one way to identify which tools have the greatest number of the important features we have identified.

*Title/Abstract Screening and Full-text Review*

These are the two basic features a screening tool should address to meet the PRISMA standard. All tools analyzed in this paper provide a mechanism for screening based on the content of the title and abstract; however, some tools provide the additional capability of streaming the included studies into the second stage of full-text review. Depending on the research goals and methods, sometimes it is sufficient to use tools that only provides title and abstract screening. For comprehensive systematic reviews, it is important for tools to handle full-text review as well since it helps generate the flow chart diagram, keeps the record of screening decisions together, and makes it more flexible to change screening decisions. This can assist researchers to better document decisions at different stages of the review process.

*Bulk Import and Export*

After literature search, researchers usually download files with their search results directly from databases or store retrieved articles into citation management tools. In both cases, citations of retrieved articles can be stored in common bibliographic formats (e.g., RIS, BibTeX, CSV). These standard data formats are compatible with a wide range of applications, including citation management tools, such as EndNote and Zotero, and screening tools, such as Covidence and Abstrackr. Bulk import and export mean two functions: First, researchers can import thousands of studies into the screening tool in their bibliographic formats; Second, upon completion of screening, researchers can export screening decisions in bulk to facilitate further coding analysis. This feature is coded 1 if the tool allows researchers to import or export thousands of studies in one single action. Bulk import and export is a necessary function for researchers to manage numerous studies in an efficient way. These features can save time for researchers, because they can efficiently combine results from multiple sources.

*Machine Learning Classifiers and Deep Learning*

Machine learning algorithms require labelled training data (such as the abstract of a study and a label that classifies the study as relevant or irrelevant) to make predictions for unseen data. In the case of systematic review, labels for data are not available a-priori. Therefore, systematic review can benefit from active learning algorithms, a special case of machine learning, which iteratively choose data points they would like the user to label in order to gradually improve classification performance (Singh et al., 2018). Machine learning has immense potential to change how modern systematic review is conducted. Traditionally, screening has consumed significant time and labor and is prone to human error. Machine learning-based text classification algorithms can help reduce time, labor, and human error by classifying or ranking studies by their relevance to expedite the systematic review process (Cohen et al., 2006). While these algorithms are increasingly being applied in the abstract and title screening stage, they are not yet ready to replace human reviewers in the full-text article quality assessment stage (Pigott & Polanin, 2020)*.*

Feature extraction algorithms process the title and abstract text and convert them into a form which can be used by text classifiers to predict the relevance of the study. Machine learning algorithms use these features to classify the text. Shallow support vector machine (SVM) models are among the most popular models used for text classification (van de Schoot et al., 2021). However, these models are increasingly being replaced by more advanced algorithms. In the last few years, deep neural network models, which use multiple non-linear processing layers to classify data, have rapidly become the state of the art due their high modeling capacity, generalizability, and performance (Young et al., 2018). For small review projects, machine learning and deep learning may not make substantial differences. For large review projects, these algorithms can expedite the screening process and tools that embed deep learning make more reliable prediction on inclusion and exclusion than tools that embed machine learning. We code the machine learning feature as 1 or 0 to differentiate whether the tools embed any text-classification machine learning algorithms. Among tools that embed machine learning algorithms, we code deep learning feature as 1 or 0 to differentiate whether the tools embed any deep neural network models.

*Teamwork and Blind Review*

For collaborative systematic reviews, web-based screening tools can support teamwork functions while desktop or terminal based tools are generally limited to local computer use. This feature is coded 1 if the tool has a web-based version that allow users to collaborate on the same project or review. In some cases, different account users are allowed to collaborate on projects only if they pay additional fees. Teamwork is an essential element for high-quality and rigorous systematic review. Many guidelines and handbooks (Cooper et al., 2019; Pigott & Polanin, 2020) list teamwork as one of the best practices for rigorous systematic reviews. Authors of these guidelines and handbooks also advocate or even require that two reviewers should screen articles independently and conduct regular team meeting to discuss and resolve conflicts (Cooper et al., 2019; Pigott & Polanin, 2020). These features can simplify the process of collaboration and independent review.

*Deduplication*

Although the function of duplication removal is available in most citation management software tools, having this function embedded in screening tools directly can save reviewers enormous time and effort. In addition, deduplication helps authors keep a systematic record of screening procedures and track the screening history for any clarifications or modifications. Most importantly, automatic deduplication prevents double counting of studies and double screening in research synthesis and prevents reviewers from wasting time screening multiple entries of the same study. Furthermore, reporting the deduplication step in research papers also assists the open science movement.

On a side note, users should be aware of the varying effectiveness in the accuracy of deduplication among screening software. Perhaps since the development and implementation of advanced tools is a recent phenomenon, there is a paucity of research that investigates and compares sensitivity, specificity, negative predictive value, and positive predictive value among screening software. One study found that Rayyan has higher accuracy and sensitivity than Covidence, but the latter has higher specificity than the former (McKeown & Mir, 2021). Another study that compared review tools with reference management tools founds that Rayyan, Mendeley, and Systematic Review Accelerator are more reliable than EndNote and Zotero (Guimarães et al., 2022). In the future, more research is required for us to gain a better understanding of the reliability of deduplication functions among screening software.

*Inter-rater Reliability*

Inter-rater reliability (IRR) is defined as the agreement between two independent coders[[1]](#footnote-2) on inclusion or exclusion decisions. IRR can be calculated at both the title and abstract screening stage as well as the full-text review stage. Reporting IRR details contributes to the transparency of systematic reviews. Although IRR can be calculated manually, enabling this function with a click of a button in the screening software could save reviewers a tremendous amount of time and effort. When the review team consists of multiple reviewers with different experience levels, IRR is an important metric to control for the review quality and reviewers’ reliability since it computes the amount of error the review intends to tolerate. Disagreements on screening decisions may arise due to different understandings of the eligibility criteria or reviewers’ physical and mental fatigue and subsequent human errors (Belur et al., 2021). To tackle the problem of “coder drift”, Polanin et al. (2019) recommends arranging regular review meetings to reconcile disagreement to reduce interpretation inconsistencies. We think that for larger review projects, using in-the-progress IRR can assist the reconciliation process as early as possible.

*Research Update*

Evidence-based reform in education has contributed to a proliferation of well-designed experiments in recent years, which constantly calls for updated systematic reviews. On average, the median update time for a systematic review is more than five years (Bashir et al., 2018). Updating systematic reviews and including the latest evidence on the topic has practical influence in evidence-based policy-making. Therefore, research update is another feature we compare in our feature analysis. We coded this feature as 1 if the tool retains the review materials and allows researchers to access review materials in the future for updating purposes.

*Reason Label*

This is a function in full-text screening to help reviewers note their reasons for making inclusion or exclusion decisions. Reason labelling is important for reviewers, as it helps them understand frequently occurring limitations in the research they are screening. In some cases, reviewers adjust either relax inclusion criteria or tighten exclusion criteria depending on whether too few or too many studies are qualified. In this scenario, reason labelling enables reviewers to re-screen excluded studies by category and adjust in a more efficient and organized way.

*Cost*

Software prices differ substantially based on the purpose (e.g., academic or business), user (e.g., student or faculty), and required functions (e.g., number of projects, number of collaborators, access to partial or all functions). Due to this variety, we code this feature qualitatively. Cost is one of the most important features because researchers’ decisions to select a screening tool can largely depend on whether they can afford it with research funding or whether access to the software is provided by the researchers’ institutions.

*Privacy Policy*

Information on privacy policy is collected from software companies’ policy documents. This feature is difficult to compare directly. We analyzed privacy policy adapting from the table of comparison developed by Bischoff (2017) when the author compared internet companies’ privacy policies. In addition to the points Bischoff (2017) coded, we added regular review of privacy policy, option to delete personal information, contact you for opinions, collects payment details, and collects device information.

**Results**

*Objective 1. Identifying available screening tools*

The web-based wide search identified 26 available screening tools in the market. Among the 26 identified tools, eight have been used by educational researchers, including Abstrackr (Wallace et al., 2012), Covidence (2013), ASReview (Schoot et al., 2021), RevMan (Review Manager, 2014), Rayyan (Ouzzani et al., 2016), EPPI-Reviewer (Thomas et al., 2020), DistillerSR (2021), and Excel spreadsheet. Table 1 presents detailed basic information about the above-mentioned eight included screening tools and other four tools not used in education yet but have the potential to be utilized. Table 2 presents features of the other 14 tools and reasons for excluding them.

*Objective 2. Assessing the tools educational researchers are already using through tool reporting in published reviews*

Article retrieval returned 191 full-text studies from *RER.* Figure 1 is a PRISMA diagram that shows the record selection process conducted in Covidence (2013). In the first full-text stage, 166 studies were excluded for not reporting any tool usage. Only 25 studies reported the use of any tools. In the second full-text stage, we excluded 12 studies for reporting tools not used for screening or citation management purposes. These 12 studies reported tools used to aid in coding information from studies, which include CMA, Nvivo, SAS, SPSS, Google forms, and Excel. In addition, we excluded five studies that reported the use of citation management tools to conduct screening, with one study using Scopus (Schrijvers et al., 2019), three studies using EndNote X7 (Braunack-Mayer et al., 2020; Kyndt et al., 2016; Surr et al., 2017), and one study using Zotero (Sabey et al., 2017).

Our research found that only eight (4.19%) studies reported tools used during screening, with one study using DistillerAI (Noetel et al., 2021), one study using EPPI-Reviewer 4 (Merrill, 2021), one study using Covidence (Lee et al., 2021), one study using Abstrackr (Bae et al., 2021), and others using Excel spreadsheets (Billingsley & Bettini, 2019; Firestone et al., 2020; Hallinger & Kovačević, 2019; Rowan et al., 2021). Figure 2 presents bar plots of the above findings.

*Objective 3. Analyzing differing features of available screening tools to inform educational meta-analysts’ tool selection*

Figure 3 presents the feature analysis score of included screening tools. The ranking from high to low for the eight tools currently used in educational research is: Covidence (1), DistillerSR (2, tied), EPPI-Reviewer (2, tied), Rayyan (4), Abstrackr (5, tied), RevMan (5, tied), ASReview (7), and Excel (8). If we include the four tools not used in education yet but have the potential to be utilized, the overall ranking becomes: Covidence (1), DistillerSR (2, tied), EPPI-Reviewer (2, tied), CADIMA (4), Swift-Active (5), Rayyan (6, tied), SysRev (6, tied), Abstrackr (8, tied), ReLiS (8, tied), RevMan (8, tied), ASReview (11), and Excel (12). Table 3 presents a detailed comparison of scores among these tools. The supplementary materials provide more information (e.g., each tool’s features, strengths, and limitations) on the eight included screening tools. The following text provides some additional comments on the feature comparison.

More than five screening tools use machine learning algorithms for title and abstract screening. All the tools we include use shallow support vector machine (SVM) models for text classification. Among the tools we included, only ASReview has the option to use deep neural network models for both classification and feature extraction, in addition to a wide range of shallow machine learning algorithms. One thing to note is that ASReview’s use of a training model returns a relevancy ranking, which means that the reviewers need to arbitrarily decide a stopping point and the number of relevant studies to extract from the list of relevancy ranking.

In terms of real-time collaboration, all included tools, except ASReview, are web-based. It is difficult to enable real-time collaboration on ASReview and almost impossible to keep track of screening decisions. ASReview does allow researchers to set up a common server using certain codes in the terminal, but the procedure is not straightforward and often requires assistance from IT department. As not all researchers have the knowledge or resources to set up custom IT infrastructure, ASReview might preclude the possibility of teamwork. Unlike researchers from STEM disciplines, educational researchers often are not required to have advanced training in Python and other programming languages. Since ASReview is Python- and terminal-based, it poses technical obstacles for systematic reviewers in the field of education.

In terms of deduplication, Covidence, Rayyan, EPPI-Reviewer, and DistillerSR enable deduplication at the title and abstract screening stage. However, not all deduplication algorithms are reliable. A previous study comparing Covidence and Rayyan (Kellermeyer et al., 2018) reports that Rayyan missed several known duplicates and thus recommends reviewers to carry out the deduplication process using citation software instead of Rayyan. For tool that cannot automatically detect or remove duplicates, reviewers need to first remove duplicates in another software (e.g., citation management software) before uploading them into these screening tools.

In terms of the cost of the software license, not all tools are free of charge. In fact, some tools are quite expensive. Only Abstrackr and ASReview are completely free of charge at the moment. Rayyan provide a free license with limited access to functions. At this time, DistillerSR is the most expensive software among all included screening tools in this paper. Since prices change frequently, readers should be aware that data on pricing were access on March 14th, 2023. Details on pricing can be found in supplementary information.

Table 4 compares tools in terms of privacy policies. Unlike many other tools, Abstrackr promises to never share users’ personal information with any third parties. Users also have the option to erase all recorded information by emailing the Brown University maintenance team. ASReview has the most secure privacy policy, which states that the software never tries to access user information.

**Discussion**

In total, this paper identifies 26 available tools to support screening large numbers of studies for systematic reviews. Not all 26 tools were suitable for the field of education. In Objective 1b, we identified 12 tools: eight screening tools used by educational researchers and four tools that are not yet utilized but have the potential for educational researchers. In our second objective, we screened published articles in *RER* and found a low rate of transparent tool reporting. In the third objective, we performed feature analysis on the 12 tools identified in Objective 1b. Based on the incorporation of the features documented in this study, these tools are more appropriate than spreadsheet or citation management tools and support higher quality research syntheses. Thus, research teams should make a tool decision that fits their specific set of needs and research questions.

Although the rapid development of machine learning algorithms provides efficient ways to classify and rank the relevance of studies, algorithms are not perfect, and researchers need to be aware of the limitations. One such limitation is the potential existence of bias in machine learning. Mehrabi et al. (2022) presented a comprehensive list of different biases in machine learning. Among these biases, representation bias, which arises when the set of data being trained does not follow the population distribution (Farquhar et al., 2021), is worth noting by systematic reviewers. In systematic review tools, supervised machine learning algorithms label unclassified studies based on a small set of studies manually classified by researchers, which is prone to representation bias. A case study on active learning in automated text classification provides evidence for the existence of such bias due to skewed sampling (Varghese et al., 2019). It is essential for tool developers to incorporate cutting-edge methods, such as novel corrective weights (Farquhar et al., 2021), to remove or compensate such biases. However, since this is still an emerging and developing field, existing screening tools may take some time to update their algorithms. One study ranked the performance of machine learning in descending order among three tools as: Abstrackr, RobotAnalyst, and DistillerSR, but the performance depends on tool and review topics (Gates et al., 2019). In the future, more research on correcting machine learning bias and comparing such bias among various tools can help with tool development and selection.

Given the limitations in machine learning, there are two ways to exercise caution and reduce algorithmic bias. One method researcher can adopt is to double check the studies excluded automatically by machine learning classifiers. For example, Zhang et al. (2022) manually scanned 10% of the studies excluded by ASReview’s machine learning algorithm to verify that these studies were indeed irrelevant. Another method is to replace one of the two human screeners by the machine screener to increase reliability since both human and machine screeners are imperfect (Gates et al., 2018). For human reviewers, an empirical study found 10.76% total error rate for false inclusion and exclusion (Wang et al., 2020). With aid from machine learning, human reviewers can focus their efforts on a smaller portion of studies and could reduce their error rate. However, for machines, there is a trade-off between saving tons of workloads and the potential of missing relevant studies. In general, we are not ready to completely automate the screening process at the moment, and semi-automation with thoughtful design is a more careful approach.

*Implications for Practice*

When making the decision on the right screening tools, researchers should first decide on the screening tool functions necessary for their particular set of study needs. This will require determining whether the studies will include quantitative or qualitative studies. The identified tools may be suitable for both types of research studies, particularly at the title and abstract screening phase, though some researchers may prefer to use qualitative software for the full-text review. Furthermore, researchers should decide on the number of reviewers on their screening team and their geographical locations to proceed with either individual research or team collaboration. Once these decisions are made, some of the tools described above can be excluded, making tool selection easier.

To assist reviewers’ tool selection process, we present a decision tree (Figure 4) based on our opinions. The first question in the decision tree asks whether reviewers are looking for a tool to review studies’ full text. The second question in the decision tree asks researchers to check their research funding and institutional resource support to ensure that the cost of the software tools is covered. Depending on the pool size of literature, time frame, and labor resource, researchers may decide whether machine learning classifiers or deep learning is necessary. Therefore, the third and forth questions ask whether reviewers are looking for a tool that has machine learning or deep learning support. This decision tree provides one logical flow to select tools. Reviewers may reverse the question order or skip non-essential questions to make informed tool-selecting decisions for their own review team. We understand that there may be other questions of interest, such as the ability to access the tool online (such as certain cloud-based products being banned in mainland China or limited Internet bandwidth in some areas making web-based tools unusable) or the ability of the tool to handle non-Latin characters. This decision tree is not meant to be comprehensive, rather to provide a starting point. Once researchers have identified a few potential tools, they can explore those in detail to determine the fit with their project goals.

Finally, while it is essential to have good tools, it is also essential that the tools should be used in the right way. After choosing a suitable tool, it is important to train the team to use the tool correctly and systematically.

*Directions for Future Research*

One problem this paper identifies is that publications seldom report the screening tools they use. Not reporting software tools for the screening procedure means that either the author(s) did not use any software, or they did not report it. If they did not use any assistive tools and culled thousands of papers manually, these researchers could benefit from leveraging existing technology for a less costly and more efficient literature screening process. The latter explanation is probably more common and reasonable, which means that the problem of scientific opacity is haunting the field of educational systematic reviews. This contradicts the PRISMA statement. The PRISMA statement illustrates the importance of reporting by stating that coherent, lucid, and transparent reporting is an important determinant of the value of a systematic review (Moher et al., 2009). Reporting the screening process in as much detail as possible is essential for future updates, reproducible research, assessment of the search’s quality, and display of scientific rigor. The current use of screening tool reporting is far from ideal, and more effort is necessary to expedite the adoption of open science practices in the field of systematic reviews.

Following the open science movement’s call for open materials and open data, researchers should try to report or preregister their systematic reviews in as much detail as possible. The choice of whether to use semi-automation in the screening stage could have a great impact on the cost-effectiveness of the research procedure and the accuracy of the research outcomes. Having the information on the specific screening tools used in each research study could better help us compare the accuracy and reliability of the tools. Future studies can benchmark the machine learning algorithms against human reviewers to compare their performance against each other on the same dataset. Researchers could quantify labor and the software license as the cost formula and performance as the benefit to calculate cost-benefit ratio for each tool. This cost-benefit ratio analysis and accuracy comparison could further provide knowledge on tool selection and its impact on systematic review. Apart from reporting the tools used, researchers should also strive to make screening files public and shareable. For instance, attaching a link to screening files in the supplementary section of publications is a common practice to promote open science. Sharing in research facilitates new scientific advances, saves duplicative effort, and helps the field to progress together.

In conclusion, this study contributes to the field of systematic reviews by comparing common screening tools used in educational research and providing practical guidance in tool selection. We advocate for researchers to select tools based on the suitability with research questions, domains, and designs, instead of on convenience of accessing tools. Moreover, we found that published systematic reviews seldomly report tools used to assist research, which may create barrier for replication. We encourage journal reviewers as well as authors to adopt open science approach to transparently report tool usage.

We have no conflicts of interest to disclose.

The authors acknowledge Hunter Gehlbach for constructive suggestions on the final draft. We also want to thank Akash Pallath and Zhipeng Hou for useful discussions on machine learning, and Susan Davis, Lisa Nehring, Rohan Arcot, Katherine Cornwall for their suggestions on the earlier draft.

**References**

Bae, C. L., Mills, D. C., Zhang, F., Sealy, M., Cabrera, L., & Sea, M. (2021). A systematic review of science discourse in k–12 urban classrooms in the united states: Accounting for individual, collective, and contextual factors. *Review of Educational Research*, 003465432110424. https://doi.org/10.3102/00346543211042415

Bashir, R., Surian, D., & Dunn, A. G. (2018). Time-to-update of systematic reviews relative to the availability of new evidence. *Systematic Reviews*, *7*(1), 195. https://doi.org/10.1186/s13643-018-0856-9

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, *50*(2), 837–865. https://doi.org/10.1177/0049124118799372

Bigendako, & Syriani. (2018). Modeling a tool for conducting systematic reviews iteratively. *Proceedings of the 6th International Conference on Model-Driven Engineering and Software Development*, 552–559.

Billingsley, B., & Bettini, E. (2019). Special education teacher attrition and retention: A review of the literature. *Review of Educational Research*, *89*(5), 697–744. https://doi.org/10.3102/0034654319862495

Bischoff, P. (2017, March 20). Comparing the privacy policy of internet giants side-by-side. *Comparitech.* https://www.comparitech.com/blog/vpn-privacy/we-compared-the-privacy-policies-of-internet-giants-side-by-side/

Bozada, T., Borden, J., Workman, J., Del Cid, M., Malinowski, J., & Luechtefeld, T. (2021). Sysrev: A FAIR platform for data curation and systematic evidence review. *Frontiers in Artificial Intelligence*, *4*. <https://www.frontiersin.org/articles/10.3389/frai.2021.685298>

Bradburn, S. (2018, July 2). 13 Best Free Meta-Analysis Software To Use. *Top Tip Bio*. https://toptipbio.com/free-meta-analysis-software/

Braunack-Mayer, A. J., Street, J. M., Tooher, R., Feng, X., & Scharling-Gamba, K. (2020). Student and staff perspectives on the use of big data in the tertiary education sector: A scoping review and reflection on the ethical issues. *Review of Educational Research*, *90*(6), 788–823. https://doi.org/10.3102/0034654320960213

Car, J., Carlstedt-Duke, J., Car, L. T., Posadzki, P., Whiting, P., Zary, N., Atun, R., Majeed, A., Campbell, J., & Collaboration, D. H. E. (2019). Digital Education in Health Professions: The Need for Overarching Evidence Synthesis. *Journal of Medical Internet Research*, *21*(2), e12913. https://doi.org/10.2196/12913

Center for Evidence Synthesis in Health. (2021). *Software | Center for Evidence Synthesis in Health | Brown University*. https://www.brown.edu/public-health/cesh/resources/software

Cheng, S. H., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minnich, R. C., Garside, R., Masuda, Y. J., Miller, D. C., Wilkie, D., Wongbusarakum, S., & McKinnon, M. C. (2018). Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation Biology*, *32*(4), 762–764. https://doi.org/10.1111/cobi.13117

Clinton, V., & Khan, S. (2019). Efficacy of Open Textbook Adoption on Learning Performance and Course Withdrawal Rates: A Meta-Analysis. *AERA Open*, *5*(3), 2332858419872212. https://doi.org/10.1177/2332858419872212

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association : JAMIA*, *13*(2), 206–219. https://doi.org/10.1197/jamia.M1929

Cooper, H., Hedges, L., & Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis* (3rd ed.). Russell Sage Foundation. <https://www.russellsage.org/publications/handbook-research-synthesis-and-meta-analysis-second-edition>

Covidence systematic review software. (2013). *Veritas Health Innovation*. www.covidence.org

DistillerSR. (2021). *Evidence Partners* (2.35) [Computer software]. https://www.evidencepartners.com

Farquhar, S., Gal, Y., & Rainforth, T. (2021). *On statistical bias in active learning: How and when to fix it* (arXiv:2101.11665). arXiv. <http://arxiv.org/abs/2101.11665>

Firestone, A. R., Cruz, R. A., & Rodl, J. E. (2020). Teacher study groups: An integrative literature synthesis. *Review of Educational Research*, *90*(5), 675–709. https://doi.org/10.3102/0034654320938128

Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews*, *7*(1), 45. <https://doi.org/10.1186/s13643-018-0707-8>

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*, *8*(1), 278. <https://doi.org/10.1186/s13643-019-1222-2>

Guimarães, N. S., Ferreira, A. J. F., Ribeiro Silva, R. de C., de Paula, A. A., Lisboa, C. S., Magno, L., Ichiara, M. Y., & Barreto, M. L. (2022). Deduplicating records in systematic reviews: There are free, accurate automated ways to do so. Journal of *Clinical Epidemiology, 152*, 110–115. https://doi.org/10.1016/j.jclinepi.2022.10.009

Hallinger, P., & Kovačević, J. (2019). A bibliometric review of research on educational administration: Science mapping the literature, 1960 to 2018. *Review of Educational Research*, *89*(3), 335–369. https://doi.org/10.3102/0034654319830380

Harrison, H., Griffin, S. J., Kuhn, I., & Usher-Smith, J. A. (2020). Software tools to support title and abstract screening for systematic reviews in healthcare: An evaluation. *BMC Medical Research Methodology*, *20*(1), 7. https://doi.org/10.1186/s12874-020-0897-3

Kellermeyer, L., Harnke, B., & Knight, S. (2018). Covidence and Rayyan. *Journal of the Medical Library Association : JMLA*, *106*(4), 580–583. https://doi.org/10.5195/jmla.2018.513

Kitchenham, B. (1996). Desmet: A method for evaluating software engineering methods and tools [Technical Report TR96-09]. University of Keele. https://silo.tips/download/desmet-a-method-for-evaluating-software-engineering-methods-and-tools

Kitchenham, B., & Linkman, S. (2000). *DESMET: A method for evaluating Software Engineering methods and tools*. https://www.semanticscholar.org/paper/DESMET-%3A-A-method-for-evaluating-Software-methods-Kitchenham-Linkman/19026a49483bef1b0a68e53743b6e53f4e7a403c

Kohl, C., McIntosh, E. J., Unger, S., Haddaway, N. R., Kecke, S., Schiemann, J., & Wilhelm, R. (2018). Online tools supporting the conduct and reporting of systematic reviews and systematic maps: A case study on CADIMA and review of existing tools. *Environmental Evidence*, *7*(1), 8. <https://doi.org/10.1186/s13750-018-0115-5>

Kyndt, E., Gijbels, D., Grosemans, I., & Donche, V. (2016). Teachers’ everyday professional development: Mapping informal learning activities, antecedents, and learning outcomes. *Review of Educational Research*, *86*(4), 1111–1150. https://doi.org/10.3102/0034654315627864

Lee, J., Sanders, T., Antczak, D., Parker, R., Noetel, M., Parker, P., & Lonsdale, C. (2021). Influences on user engagement in online professional learning: A narrative synthesis and meta-analysis. *Review of Educational Research*, 003465432199791. https://doi.org/10.3102/0034654321997918

McKeown, S., & Mir, Z. M. (2021). Considerations for conducting systematic reviews: Evaluating the performance of different methods for de-duplicating references. *Systematic Reviews, 10*(1), 38. https://doi.org/10.1186/s13643-021-01583-y

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). *A survey on bias and fairness in machine learning* (arXiv:1908.09635). arXiv. <http://arxiv.org/abs/1908.09635>

Merrill, B. C. (2021). Configuring a construct definition of teacher working conditions in the united states: A systematic narrative review of researcher concepts. *Review of Educational Research*, *91*(2), 163–203. https://doi.org/10.3102/0034654320985611

Microsoft. (2021). *Microsoft Excel Spreadsheet Software | Microsoft 365.* <https://www.microsoft.com/en-us/microsoft-365/excel>

Miller K, Howard BE, Phillips J, Shah MR, Mav D, Shah RR (2016). “SWIFT-Active Screener: Reducing literature screening effort through machine learning for systematic reviews.” Poster presentation at the Society of Toxicology’s 55th Annual Meeting and ToxExpo, New Orleans, LA.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1136/bmj.b2535

Noetel, M., Griffith, S., Delaney, O., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2021). Video improves learning in higher education: A systematic review. *Review of Educational Research*, *91*(2), 204–236. https://doi.org/10.3102/0034654321990713

Ødegaard, N. B., Myrhaug, H. T., Dahl-Michelsen, T., & Røe, Y. (2021). Digital learning designs in physiotherapy education: A systematic review and meta-analysis. *BMC Medical Education*, *21*(1), 48. https://doi.org/10.1186/s12909-020-02483-w

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Pallath, A., & Zhang, Q. (2022). Paperfetcher: A tool to automate handsearching and citation searching for systematic reviews. *Research Synthesis Methods*, 1(13). 1- 13. doi:10.1002/jrsm.1604

Patall, E. A. (2021). Implications of the open science era for educational psychology research syntheses. *Educational Psychologist*, *56*(2), 142–160. <https://doi.org/10.1080/00461520.2021.1897009>

Pei, L., & Wu, H. (2019). Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. *Medical Education Online*, *24*(1), 1666538. https://doi.org/10.1080/10872981.2019.1666538

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods, 10*(3), 330–342. https://doi.org/10.1002/jrsm.1354

Polanin, J. R., Espelage, D. L., Grotpeter, J. K., Spinney, E., Ingram, K. M., Valido, A., El Sheikh, A., Torgal, C., & Robinson, L. (2021). A meta-analysis of longitudinal partial correlations between school violence and mental health, school performance, and criminal or delinquent acts. *Psychological Bulletin*, *147*(2), 115–133. https://doi.org/10.1037/bul0000314

Rathbone, J., Hoffmann, T., & Glasziou, P. (2015). Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews, 4*(1), 80. https://doi.org/10.1186/s13643-015-0067-6

Review Manager. (2014). *RevMan[Computer program]* (5.3) [Computer software]. The Cochrane Collaboration.

Roth, S. (2021). *Research Guides: Systematic Reviews & Other Review Types: Systematic Review Tools*. https://guides.temple.edu/systematicreviews/SRTools

Rowan, L., Bourke, T., L’Estrange, L., Lunn Brownlee, J., Ryan, M., Walker, S., & Churchward, P. (2021). How does initial teacher education research frame the challenge of preparing future teachers for student diversity in schools? A systematic review of literature. *Review of Educational Research*, *91*(1), 112–158. https://doi.org/10.3102/0034654320979171

Sabey, C. V., Charlton, C. T., Pyle, D., Lignugaris-Kraft, B., & Ross, S. W. (2017). A review of classwide or universal social, emotional, behavioral programs for students in kindergarten. *Review of Educational Research*, *87*(3), 512–543. https://doi.org/10.3102/0034654316689307

Salter, S. M., Karia, A., Sanfilippo, F. M., & Clifford, R. M. (2014). Effectiveness of e-learning in pharmacy education. *American Journal of Pharmaceutical Education*, *78*(4), 83. https://doi.org/10.5688/ajpe78483

Schoot, R. van de, Bruin, J. de, Schram, R., Zahedi, P., Boer, J. de, Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

Schrijvers, M., Janssen, T., Fialho, O., & Rijlaarsdam, G. (2019). Gaining insight into human nature: A review of literature classroom intervention studies. *Review of Educational Research*, *89*(1), 3–45. https://doi.org/10.3102/0034654318812914

Singh, G., Thomas, J., & Shawe-Taylor, J. (2018). Improving active learning in systematic reviews. *ArXiv:1801.09496 [Cs]*. http://arxiv.org/abs/1801.09496

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11.

Storey, Neitzel, A., Wolf, R., Guo, X., Slavin, R. E., Madden, N., & Collins, M. (2021). A randomized clinical trial evaluating learning impacts of provision of eyeglasses to students in grades prek-2. *SREE2021*. Society for Research on Educational Effectiveness.

Surr, C. A., Gates, C., Irving, D., Oyebode, J., Smith, S. J., Parveen, S., Drury, M., & Dennison, A. (2017). Effective dementia education and training for the health and social care workforce: A systematic review of the literature. *Review of Educational Research*, *87*(5), 966–1002. https://doi.org/10.3102/0034654317723305

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O’Driscoll, P., & Bond, M. (2020). *EPPI-Reviewer: Advanced software for systematic reviews, maps and evidence synthesis.* EPPI-Centre Software. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2967

Van der Mierden, S., Tsaioun, K., Bleich, A., & Leenaars, C. H. C. (2019). Software tools for literature screening in systematic reviews in biomedical research. *Altex*, *36*(3), 508–517. https://doi.org/10.14573/altex.1902131

Varghese, A., Hong, T., Hunter, C., Agyeman-Badu, G., & Cawley, M. (2019). Active learning in automated text classification: A case study exploring bias in predicted model performance metrics. *Environment Systems and Decisions*, *39*(3), 269–280. <https://doi.org/10.1007/s10669-019-09717-3>

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 819–824. https://doi.org/10.1145/2110363.2110464

Wang, J., & Leeflang, M. (2019). Recommended software/packages for meta-analysis of diagnostic accuracy. *Journal of Laboratory and Precision Medicine*, *4*(0). https://doi.org/10.21037/jlpm.2019.06.01

Wang, Z., Nayfeh, T., Tetzlaff, J., O’Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PLOS ONE*, *15*(1), e0227742. <https://doi.org/10.1371/journal.pone.0227742>

Westgate, M. J. (2019). revtools: An R package to support article screening for evidence synthesis. *Research Synthesis Methods*, *10*(4), 606–614. https://doi.org/10.1002/jrsm.1374

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ArXiv:1708.02709 [Cs]*. http://arxiv.org/abs/1708.02709

Zhang, Q., Grant, A., Pellegrini, M., & Neitzel A. (2021). A meta-analysis of teacher salary and turnover in the United States and China. *AERA2021*. Annual meeting of the American Educational Research Association, Virtual.

Zhang, Q., Wang, J. & Neitzel, A. (2022). School-based Mental Health Interventions Targeting Depression or Anxiety: A Meta-analysis of Rigorous Randomized Controlled Trials for School-aged Children and Adolescents. *J Youth Adolescence 52,* 195–217 (2023). <https://doi.org/10.1007/s10964-022-01684-4>

**Table 1**

*Basic Information on the Included Screening Tools*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Info** | **Year of Launch** | **Developer  Publication** | **Developer** | **Country** | **Examples in Education** |
| Abstrackr | Semi-automated tool for teamwork in meta-analysis | 2012 | Wallace et al. | Tufts Evidence-based Practice Center, maintained by Brown Center for Evidence Synthesis in Health | USA | (Clinton & Khan, 2019);  (Polanin et al., 2021) |
| Covidence | Web-based systematic review tool for screening, data extraction, and analysis | 2013 | - | Australian non-profit company Veritas Health Innovation | Australia | (Car et al., 2019);  (Lee et al., 2021) |
| ASReveiw | Uses the latest machine learning algorithms to minimize errors and maximize accuracy | 2019 | Schoot et al. | Utrecht University | Netherlands | (Zhang et al., 2022) |
| RevMan Web/RevMan 5 | -Facilitates protocol development, screening, and full-text reviews  -Present the results graphically | 2008 | The Cochrane Collaboration | The Cochrane Collaboration | UK | (Pei & Wu, 2019) |
| Rayyan | Accelerates abstract/title screening with semi-automation | 2014 | Ouzzani et al. | Qatar Computing Research Institute:  Rayyan Systems Inc. | Qatar | (Ødegaard et al., 2021) |
| EPPI-Reviewer | -Supports  study screening through data collection, analysis and synthesis  -Includes features such as text mining, data clustering, classification, term extraction, and machine learning | 2010 | Thomas et al. | EPPI-Centre at the Social Science Research Unit at the Institute of Education, University College London, and University of London | UK | (Merrill, 2021) |
| DistillerSR | Automates literature collection,  triage, and assessment using AI and intelligent workflows | 2018 | - | Evidence Partners Inc. Ottawa | Canada | (Salter et al., 2014);  (Noetel et al., 2021) |
| Excel | Excel's VonVille can help with systematic reviews | 1985 | - | Microsoft Corporation | USA | (Billingsley & Bettini, 2019); (Firestone et al., 2020); (Hallinger & Kovačević, 2019); (Rowan et al., 2021) |
| **Other four tools not used in education yet but have the potential to be utilized** | | | | | |  |
| SysRev | SysRev helps systematic reviewers with machines to extract data from documents. It claims to be built for general purpose data miners and is available on github (<https://github.com/sysrev/Sysrev_Documentation> ) | 2021 | (Bozada Jr et al., 2021) | Insilica | USA | - |
| SWIFT-Active Screener | SWIFT-Active Screener is a web-based, collaborative systematic review software application. It was developed by Sciome LLC, which is an innovative research and technology consulting company. | 2016 | (Miller et al., 2016) | Sciome LLC | USA | - |
| CADIMA | CADIMA is a free web tool facilitating the conduct and assuring  for the documentation of systematic reviews, systematic maps and further literature reviews. | 2018 | (Kohl et al., 2018) | Julius Kühn-Institut | Germany | - |
| ReLiS | ReLiS stands for Revue Litteraire Systématique which is French for Systematic Literature Reviews ReliS literally translates to “reread”. It is available on github (<https://github.com/geodes-sms/relis> ). ReLiS is a highly configurable tool to conduct systematic reviews collaboratively and iteratively on the cloud. | 2018 | (Bigendako & Syriani, 2018) | Software engineering lab [GEODES](http://geodes.iro.umontreal.ca/en/) in the [department of computer science and operations research](http://diro.umontreal.ca/accueil/), University of Montreal | Canada | - |

**Table 2**

*Other 14 Screening Tools*

|  |  |  |
| --- | --- | --- |
| **Name** | **Reason for exclusion** | **Information** |
| Colandr | Biodiversity | Colandr is an open-source and open-access machine-learning assisted online platform for  conducting systematic reviews and syntheses with a focus in biodiversity research (Cheng et al., 2018). |
| SyRF | Preclinical | SyRF is a fully integrated online platform for conducting systematic reviews of  preclinical studies developed by researchers from The University of Edinburgh. It is completely free for researchers. |
| RobotAnalyst | Public health | RobotAnalyst was developed to support systematic reviews in public health interventions. It supports literature screening with machine learning and text mining algorithms. It adopts topic modelling and relevance feedback-based text classification models. |
| SRDR Plus | Healthcare | SRDR Plus (The Systematic Review Data Repository: Plus) is a free tool for extracting, managing, and archiving data. It was developed by researchers at Brown University to support healthcare research primarily. It contributes to open science through an open systematic review data repository. |
| DRAGON | Health and environmental research | DRAGON, launched in 2018, stores qualitative and quantitative data from literature to help scientists implement the elements of systematic review, including problem formulation, literature screening, risk of bias evaluation, and data integration. It was developed for health and environmental research. |
| SRDB.PRO | Pharmaceutical industry and health, and health economics consultancies | SRDB.PRO is the first enterprise level systematic review and data analysis  platform designed specifically for the pharmaceutical industry and health, and health economics consultancies. It integrates the PubMed literature database. It costs $119 for a two-user commercial licence and $76,500 for a 5-user commercial licence. For an academic license, the cost is $0 for two users and $8,500 for 100 users. |
| BioReader (Biomedical Research Article Distiller) | Biomedical research | BioReader is a tool that enables users to perform classification of scientific literature by text mining-based classification of article abstracts. The tool is trained by uploading article corpora for two training categories - e.g., one positive and one negative for content of interest - as well as one corpus of abstracts to be classified and/or a search string to query PubMed for articles. The corpora are submitted as lists of PubMed IDs and the abstracts are automatically downloaded from PubMed, preprocessed, and the unclassified corpus is classified using the best performing classification algorithm out of ten implemented algorithms. BioReader is freely available as a web service. |
| PASRIFAL | Software engineering | Parsifal is an online tool designed to support researchers to perform systematic literature reviews within the context of Software Engineering. It claims to support geographically distributed researchers. |
| Thoth | Software engineering | Thoth is a web-based support tool developed to support the SLR process in software engineering. |
| StArt | Software engineering | State of the Art through systematic review (StArt) aims to provide support for each stage of the SR process in software engineering. It was developed in Brazil by The Laboratory of Research on Software Engineering (LaPES), the Computing Department of the Federal University of São Carlos (DC/UFSCar). Available at <http://lapes.dc.ufscar.br/tools/start_tool> |
| Revtools | R package | Revtools is an R package to support article screening for evidence synthesis. It supports importing, deduplication, title/article screening, and visualization of article content using topic models (Westgate, 2019). |
| METAGEAR | R package | Metagear is a R package published in 2021. It is a comprehensive, multifunctional toolbox with capabilities aimed to cover much of the research synthesis taxonomy: from applying a systematic review approach to objectively assemble and screen the literature, to extracting data from studies, and to finally summarize and analyze these data with the statistics of meta-analysis. It is available on github (<https://github.com/mjlajeunesse/metagear> ) |
| FASTREAD | Machine learning framework | FASTREAD is one of the state-of-the-art automatic methods to expedite reference screening with study prioritization. It is open-source and open-access and is available on github (<https://github.com/fastread/SLR_on_TCP> ). It is a machine learning framework. |
| DBPedia | Framework | A resource description framework repository to support automated selection of primary studies. |

**Table 3**

*Features of the Eight Screening Tools*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Platform** | **General Functions\*\*\*** | | | | | **Title/abstract screening stage** | | | | **Full-text review stage** | | | |  |
| **Bulk  import/export** | **Team**  **work** | **Blind  review** | **De-duplicate** | **Update** | **Title/ Abstract Screening** | **ML** | **Deep learning** | **IRR** | **Full-text  review** | **IRR** | **Reason**  **label\*** | **Decision labels\*\*** | **Licensing** |
| Abstrackr | Web-based | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | Relevant,  borderline,  irrelevant | Free |
| Covidence | Web-based | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | Include,  maybe,  exclude | Not free  $240~$635 |
| ASReveiw | Terminal &  Python-based | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | Relevant,  irrelevant | Free |
| RevMan Web/RevMan 5 | Web-based or  desktop software | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | - | $72.87 ~ $120.24 and more |
| Rayyan | Web-based | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | Include, undecide, exclude | Free with limited functions, $48~$99 and more |
| EPPI-Reviewer | Web-based | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | Exclude,  include for second opinion, include | Not free  $144.54 ~ $505.89 and more |
| DistillerSR | Web-based | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | Yes, no, can't tell | Not Free  $239.4 ~ $3636 |
| Excel | Web-based or desktop software | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | - | Free |
| Sysrev | Web-based | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |  | Not Free  $0 ~ $120 and more |
| SWIFT-Active Screener | Web-based | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | Include, exclude | Not Free  Non-transparent pricing |
| CADMIA | Web-based | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | Criteria or comment | Free |
| ReLiS | Web-based | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | Include, exclude | Free |

Abbreviations: IRR., Inter-rater reliability. ML = Machine learning classifiers. \*Reason label means whether the tool allows users to customize comments on the reasons for their decisions. \*\*Decision label means labels provided by the tools when making a decision. \*\*\*General functions’ scores are only counted for full-text review score if the tool enables full-text review.

Note: When a feature is labelled 0, it means that the authors did not find the specific function, which does not necessarily mean that the tool lacks this function. In licensing prices, price range is the annual subscription fee in US dollars as per information accessed on March 14th, 2023. Detailed information on price plans is provided in the supplementary information. Raw data used to create this table is available through github: <https://github.com/qiyangzh/Data-on-Choosing-the-right-tool-for-the-job-Screening-tools-for-systematic-reviews-in-education.git>

**Table 4**

*Privacy Policy Feature Analysis*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tool** | **Adds Cookies** | **Collects Personal Information** | **Collects Device Information** | **Collects Payment Details** | **Contact You for Opinions** | **Collects Location** | **Uses 1st party ads** | **Uses 3rd party ads/links** | **Shares personal info with 3rd party** | **Stores info in multiple countries** | **Indefinite Retention of Information** | **Regular Review of Privacy Policy** | **Option to Delete Personal Information** |
| Abstrackr | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Covidence | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ASReveiw | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| RevMan | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| Rayyan | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| EPPI-Reviewer | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| DistillerSR | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Excel | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| SysRev | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Swift-Active | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| CADIMA | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| ReLiS | NA | | | | | | | | | | | | |

Note: When a feature is labelled 0, it means that the authors did not find the specific function, which does not necessarily mean that the tool lacks this function. ReLiS does not have privacy policy available online.

**Figure 1**

*PRISMA Diagram*

**Diagram

Description automatically generated**

**Figure 2**

*Barplot for Citation Management Tool and Screening Tool Reporting Status*

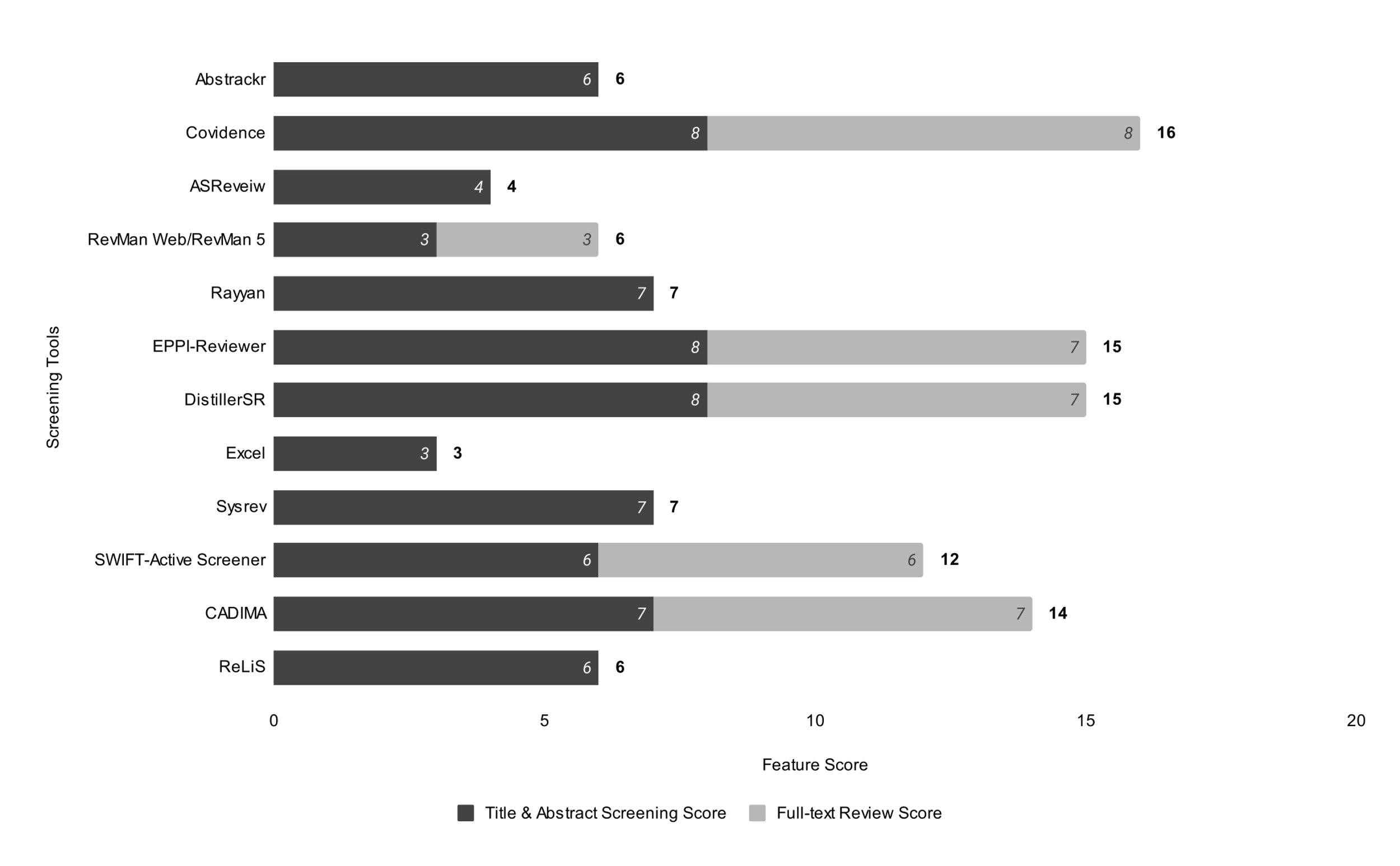
Chart

Description automatically generated

*Note.* The R code used to generate this figure is available on github: <https://github.com/qiyangzh/Data-on-Choosing-the-right-tool-for-the-job-Screening-tools-for-systematic-reviews-in-education.git>

**Figure 3**

*Feature Analysis Score Ranking*

****

**Figure 4**

*Decision Tree*

**

*Note.* The R code used to generate this figure is available on github: <https://github.com/qiyangzh/Data-on-Choosing-the-right-tool-for-the-job-Screening-tools-for-systematic-reviews-in-education.git>

1. Coders refer to researchers who document the characteristics of each included study to foster a deeper understanding of the study quality [↑](#footnote-ref-2)