Received: 30 October 2018 | Revised: 1 April 2019 | Accepted: 2 May 2019

DOI: 10.1002/jrsm.1354

TUTORIAL

WILEY Research Synthesis Methods

Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses

Abstract screening is one important aspect of conducting a high-quality and comprehensive systematic review and meta-analysis. Abstract screening allows the review team to conduct the tedious but vital first step to synthesize the extant literature: winnowing down the overwhelming amalgamation of citations discovered through research databases to the citations that should be "full-text" screened and eventually included in the review. Although it is a critical process, few guidelines have been put forth since the publications of seminal systematic review textbooks. The purpose of this paper, therefore, is to provide a practical set of best practice guidelines to help future review teams and managers. Each of the 10 proposed guidelines is explained using real-world examples or illustrations from applications. We also delineate recent experiences where a team of abstract screeners double-screened 14 923 abstracts in 89 days.

1 | INTRODUCTION

Conducting a systematic review and meta-analysis, large or small, requires dedicated planning, consistent information tracking, and constant managerial oversight.[1] A high-quality review relies on a team of content and methodological members' expertise combined with knowledge cultivated through the completion of previous reviews. A major task of a systematic review is the identification of studies eligible for a review and then the screening of these studies to find those eligible for the review. Searching and identifying a wide range of studies for a systematic review, therefore, is critical for a high-quality systematic review and meta-analysis.

A challenge for systematic review teams in the social sciences is that many research questions transcend disciplinary boundaries, requiring that the search for relevant studies includes the use of several disciplinary and cross-disciplinary databases. The social and behavioral sciences, in addition, have yet to adopt structured abstract guidelines, complicating keyword searches for eligible studies. Many researchers conducting a systematic review commonly identify thousands of potentially relevant studies in initial search strategies. Examples of database searches returning over 5000 hits can readily be found in psychology,[2] education,[3] criminal justice,[4] and medicine.[5] These large-evidence reviews require organized processes to identify eligible studies efficiently while minimizing potential bias.

The typical process of identifying eligible studies for a systematic review and meta-analysis begins with screening abstracts. Abstract screening allows the review team to winnow down the large number of identified studies to the citations that should be "full-text" screened and eventually included in the review.[6] Systematic reviews aim to identify all applicable and potentially eligible studies on a topic. After conducting a comprehensive literature search to identify studies, abstract screening is the next stage in the process where bias could threaten the validity of the identified studies: abstract screening actively eliminates studies from the review and therefore the potential for bias must not be overlooked.[7] Yet, efficient and accurate abstract screening has the potential to result in a significant reduction in review time because every citation that is correctly removed represents decreased resource burden in the full-text screening phase. Any reduction in resources

Joshua R. Polanin, PhD, Principal Researcher, American Institutes for Research, 1000 Thomas Jefferson St. NW, Washington, DC, 20007. Phone: 202-403-5509. Email: jpolanin@air.org. This project was supported by Award No. 2016-CK-BX-0012, awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice, and by Award No. R305B170019, awarded by the Institute of Education Sciences, National Center for Education Research, US Department of Education. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice or Department of Education.

needed is especially important when the size of literature search results in many thousands of citations.

Systematic review and meta-analysis researchers, therefore, must dedicate resources and planning time to think critically about how, when, and where review team members will screen abstracts. Various guidelines and standards exist for general advice about screening studies for a systematic review. The Cochrane Handbook provides a typical process for selecting studies that includes a two-step process for examining titles and abstracts and then full texts.[8] Guidelines from the Center for Reviews and Dissemination,[9] the US Institute of Medicine,[10] and the Agency for Healthcare Research and Quality[11] all provide similar advice to the Cochrane Handbook for general steps for avoiding bias in screening studies for inclusion. Classic systematic review and meta-analysis textbooks also detail the basics of the abstract screening process; yet, many textbooks[12,13] lack practical information about best practice methods for this stage such as the use of text-mining software, pilot training of coders, reconciliation, or managing screeners. Other more recent research focuses on improving the efficiency of screening abstracts by evaluating the effectiveness of text-mining software.[14-16] These studies focus on the efficiency of the algorithms used for text-mining but fail to provide practical advice on how review teams can implement text-mining software in large-evidence systematic review projects.

With these issues in mind, the goal of this paper is to provide a best practice set of guidelines (Table 1) for abstract screening that discusses the management of a large-evidence review process. We start by outlining our recent experiences conducting abstract screening for a large-evidence project. Then, we discuss 10 best practice guidelines and provide real-world examples for each. We end by summarizing the guidelines, their applications, and the potential impact small decisions have when dealing with large-evidence reviews.

2 | LARGE-EVIDENCE ABSTRACT SCREENING EXAMPLE: PREDICTING MENTAL HEALTH, CRIMINALITY, AND SCHOOL PERFORMANCE FROM SCHOOL VIOLENCE EXPOSURE

A recent practical experience has helped shape the best practice guidelines presented here. The objective of the example project is to synthesize primary studies that followed K-12 grade students across two time points, using their exposure to, perpetration of, or victimization from school violence to predict later mental health, criminality, and school performance (https://osf.io/6hak7/). We provide our screening tool, which follows the best practice guidelines outlined in Online Appendix A. Note that we wrote simple and objective items (guideline 1), included yes/unsure or no answers (guideline 1), began with easily answerable items (guideline 2), and developed the tool to be used with a text-mining abstract screening tool (guideline 6). Note also that our original screening tool included all but one question, which we added after conducting an initial pilot screening session (guideline 3). We continued to add examples throughout the screening process (guideline 5) during our weekly meetings (guideline 4). Since using this screening tool on this project, our team has conducted several additional large-evidence systematic reviews, and it was beneficial to return to documentation written during a debrief of the screening process (guideline 10).

To illustrate our screening process, Table 2 delineates various milestones achieved during our abstract screening process. Over the course of 189 days, research staff screened 29,846 abstracts independently (14,923 unique citations were double-screened). Nineteen team members screened at least 100 abstracts. The average person

TABLE 1 Abstract screening best practices guidelines

Number	Screening Stage	Guideline
1	Beginning	Create an abstract screening tool with questions that are clear and concise. It should include items that (a) are objective, (b) are "single-barreled," (c) use the same sentence structure, and (d) include yes/no/unsure answers only.
2	Beginning	Ensure that the abstract screening tool is organized hierarchically, with the easiest questions at the beginning of the tool.
3	Beginning	Conduct introductory abstract screening trainings where screeners learn and pilot test the tool by screening the same 20 to 30 abstracts. Repeat as necessary until team reaches consensus.
4	During	Meet with the abstract screening team on a weekly or every other week basis.
5	During	Minimize changes to the screening tool.
6	During	Use a text-mining abstract screening application.
7	During	Require independent double-screening of each abstract.
8	During	Reconcile disagreements throughout the abstract screening process.
9	During	Encourage screening through intellectual buy-in and incentives.
10	End	Analyze the process and decisions after screening has been completed.

TABLE 2 Abstract screening process for example study

Date	Activity	Abstracts Screened
04/01/2017	PI creates screening tool	0
04/06/2017	PI, CPI, and PD meet to pilot screening tool	0
04/10/2017	PI and RA update screening tool based on pilot	0
04/12/2017	PI and RA create PWPT for screening training	0
04/14/2017	PI creates Abstrackr project; 14 923 citations uploaded; 29 846 abstracts to double-screen	0
04/16/2017	PI, CPI, and PD lead screening training with screeners; screeners assigned the same 30 abstracts; eight students, three staff, PI, CPI, and PD	0
04/21/2017	Meeting to discuss same 30 abstracts; incentives explained; individuals who completed same 30 abstracts allowed to screen on their own	30
04/25/2017	Screening tool updated to include various new descriptors	1213
05/04/2017	Update meeting; two additional students and one staff trained	3538
05/11/2017	29% disagreement rate; PI, CPI, and PD review disagreements	6249
05/15/2017	28% disagreement rate; PI, CPI, and PD decide to conduct early reconciliation	6541
05/16/2017	PD sends each screener a spreadsheet that lists each abstract where another person has screened the same one and the other person's decision; reconciliation begins	6541
05/24/2017	Screening tool updated based on reconciliation decisions	6541
05/26/2017	Screeners continue after reconciliation; 15.2% disagreement rate	6541
06/08/2017	Update meeting; 13.5% disagreement rate	8466
06/12/2017	Update meeting; 10.1% disagreement rate	12 597
06/16/2017	Updated meeting; 9.5% disagreement rate	14 359
06/21/2017	Update meeting; 9.1% disagreement rate	17 816
06/27/2017	Update meeting; 8.4% disagreement rate	21 360
06/28/2017	Update meeting; 8.4% disagreement rate	27 128
06/29/2017	Screening complete; 8.2% disagreement rate	29 846
06/30/2017	Meeting to discuss final reconciliation; PD sends spreadsheet to each individual listing disagreements	29 846
07/12/2017	Reconciliation update: 3% disagreement rate	29 846
07/19/2017	Reconciliation complete: 0% disagreements	29 846

screened 1589 abstracts (*SD* = 1531) with a median of 1001 abstracts screened. On an average day, 335 abstracts were screened. We awarded two prizes after screening ended (guideline 9): one award to the individual with the most abstracts screened and one award to the individual with the smallest proportion of disagreements. The abstract screening process resulted in the retrieval of approximately 2000 (~13%) study PDFs.

3 | BEST PRACTICE GUIDELINES FOR ABSTRACT SCREENING LARGE-EVIDENCE REVIEWS

The guidelines presented in Table 1 represent a synthesis of textbook recommendations, discussion with experts, and practical experience conducting or participating in numerous large-evidence systematic reviews and meta-analyses. It is by no means an exhaustive or exclusive list and other review teams may develop additional guidelines. It is, moreover, not a static list because text-mining programs, in particular, will continue to advance and the best practices must advance with them.

We should also note that the guidelines in this paper may apply directly to large-evidence systematic reviews and may not result in a positive return on investment for smaller review projects. We consider large-evidence reviews those above one thousand citations found in the search process, either from online databases, gray literature searches, reference harvesting, or contacting authors. As a result of the high number of citations, the process of organizing citations and screening potential studies is a nontrivial task requiring several weeks or months of work. Although we advocate the use of text-mining software for abstract screening, the cost of learning to implement the software may be greater than the efficiency benefit for smaller review projects. Review searches that yield fewer than 300 to 500 returned citations may be better served using a reference manager or EXCEL. For larger projects, many of these best practice guidelines will result in efficiency gains. And perhaps most importantly, we expect that the size of reviews will continue to grow as the number of high-quality primary studies increases.[17]

We also believe that these guidelines fit within long-standing systematic review and meta-analysis procedures. Cooper (2009)[12] provides seven steps for a systematic review and meta-analysis project. Following Cooper's steps, the abstract screening process is conducted after step 2 (searching the literature) but before step 3 (gathering information from studies). Review teams usually conduct abstract screening by examining the titles and abstracts of the studies identified as potentially eligible from the literature search. Typically, the review team develops an abstract screening tool consisting of a set of simple eligibility criteria that are reported in the abstract of a study. The review team uses the abstract screening tool to decide whether a study identified in the search is eligible for the review. After an abstract is deemed eligible by the screeners, the research team obtains the full-text document of the study, usually in PDF form. Once all PDFs have been located, the team screens the full text of documents to verify the study's eligibility. Although most of these guidelines can be applied to full-text screening, the focus of the present work is on abstract screening.

The guidelines for abstract screening of large systematic reviews are organized below by the stage of the abstract screening process: before beginning the screening, during the screening, and at the end of the screening process. We expect that some variation in the order of their implementation will occur, especially as less experienced reviewers attempt to apply the practices. This is not a step-by-step guide to conducting abstract screening; rather, these guidelines should be used like lampposts: follow them and the path to conducting better, more efficient abstract screening should be clearer. The following describes our 10 best practice guidelines for abstract screening large-evidence reviews.

4 | BEFORE SCREENING BEGINS

1. Create an abstract screening tool with questions that are clear and concise. It should include items that (a) are objective, (b) are "single-barreled," (c) use the same sentence structure, and (d) include yes/no/unsure answers only.

This first guideline follows decades of suggestions[18] about the development of screening and coding forms for systematic review. The abstract screening tool is based on the inclusion criteria of the study guiding the review and preferably published in a research protocol created prior to the literature search.[19] The abstract screening tool guides screeners in their decisions about whether a citation is eligible for the review and thus plays an important role in identifying a representative sample of studies for the review.

Given the screening tool's importance in identifying eligible studies, abstract screening questions should not be subjective for the screener. For example, the question "Does the abstract indicate that a high-quality design was used?" is ambiguous because "high-quality" is subjective. A better question might ask "Does the study use a randomized controlled trial design?" The answer to this question can easily and quickly be ascertained from the abstract.

In addition, the abstract screening tool questions should be "single-barreled," meaning each question should ask the screener about one aspect of the abstract. The question "Does the abstract indicate that adults over 18 were sampled and that the sample was from a general population?" asks about the age of the sample as well as the sample's characteristics. If each of these aspects is important, then each should be reflected in separate screening questions.

We also suggest that the questions follow a similar sentence structure. An abstract screening tool, for example, might include two questions: (a) "Was the study published on or after 1987?" and (b) "Did the study evaluate an ADHD-symptom reduction medication?". Although the questions are unambiguous and single-barreled, the differences in sentence structure may confuse abstract screeners and make learning difficult. We suggest, therefore, that the questions follow the same structure throughout the tool. To illustrate, a reviewer might change these two questions to (a) "Was the study published on or after 1987?" and (b) "Was the study an evaluation of an ADHD-symptom reduction medication?"

The answers to each abstract screening question should follow the same format: (a) yes, (b) no, or (c) unsure. Answering yes/no/unsure allows for quick and uninterrupted abstract screening. Forcing abstract screeners to provide a detailed response decreases efficiency, particularly when screeners are examining hundreds of study titles and abstracts. Review teams should ask for more details about a study during the full-text screening and coding process, not at the abstract screening stage.

We also suggest that all "yes" and all "no" answers result in the same action (ie, contributes toward being eligible or toward being ineligible) for each question. To illustrate, assume that a reviewer is interested in studies that sampled postsecondary students only, and that a "yes" indicates that the abstract is potentially eligible for the review. Therefore, an inappropriate question is "Does the study sample students that are K-12 aged?" because, if the answer to that question is yes, it means the study is ineligible.

While we advocate for including an "unsure" option, we also encourage review managers to emphasize how "unsure" should be used. When a screener answers "unsure" to a question, the study remains eligible for full-text screening provided all other inclusion criteria are met. We encourage abstract screeners, therefore, to use "unsure" only in cases where the information is not provided in the abstract. It is far more efficient to train abstract screeners to confidently answer yes or no to each question rather than to have abstract screeners who answer "unsure" to multiple questions. Many unsure answers result in multiple dispute resolutions or in many full-text articles that need to be downloaded and screened, thus increasing the time and resources needed. We strongly encourage review managers to prevent overuse of the "unsure" code as much as underuse.

One additional note is required on the "unsure" option, particularly concerning citations missing bibliographic information or entire abstracts. The screening tool works well when bibliographic information derives from well-organized databases such as ERIC. The vast majority of the time, the complete citation is available, along with the abstract and any supplementary information provided by the database. In this case, an abstract screener will have no issues in making a screening decision. Problems sometimes arise, however, when reviewers search gray literature databases or other bibliographic databases. When this occurs, we direct our screeners to continue using the screening tool as written, with the recognition that many or most answers will be marked as "unsure." Without a definitive "no" to one of the screening questions, the review team is forced to retrieve an article that may not be applicable, slowing the review process and expending review resources unnecessarily. Although we would like to suggest an alternative to this approach, we do not yet know of an acceptable practice to combat it, short of eliminating any citation missing an abstract.

2. Ensure that the abstract screening tool is organized hierarchically, with the easiest questions at the beginning of the tool

Review team members will find screening large numbers of study abstracts a tedious task. Abstract screeners naturally attempt to speed through the process and make decisions about each abstract as quickly as possible. Their speed often corresponds to their fatigue: less fatigue, all else being equal, means quicker and more reliable abstract screening. Moving quickly, yet accurately, through many abstracts during an abstract screening process, therefore, should be encouraged through the hierarchical arrangement of abstract screening questions.

We suggest, for quick and accurate screening, that the abstract screening tool begin with the easiest screening questions and end with the most difficult questions. This allows screeners to quickly screen out abstracts that clearly fail to meet the easiest to identify inclusion criteria. Here, we mean "easy" questions as ones that can be answered without interpretation, investigation, or assumptions. A great example is a question that requires abstract screeners to read only the citation (ie, "Is the date of publication on or after 1995?") or the language of the abstract (ie, "Is the abstract written in English or French?").

Difficult questions are those that require the abstract screener to make inferences from the text or that require reading the entire abstract carefully. An example of a difficult question is one that asks about the population sampled (ie, "Does the sample include participants with a disability?"). The screener must carefully read and interpret the abstract and make a judgment based on how the authors describe the sample for the study.

Livoreil and colleagues[20] and Brunton, Stansfield, Caird, and Thomas[21] both suggest the hierarchical ordering of abstract screening questions to increase efficiency. Hierarchical ordering of questions means that if a screener says no to any question, the study is ineligible and screening can stop. For example, if the screener says no to the first question, "Is the date of the study on or after 1995?", the screener can eliminate the study and move on to the next study. In addition, if the answer to the third screening question is "yes" for inclusion in the review, then the answers to the prior two questions should also be "yes."

To be clear, we are not suggesting that screeners must answer each question in the order of the screening tool. For the screening process to truly represent an efficiency, in fact, a screener should screen out an abstract as soon as the screener can identify a definitive "no." Sometimes, it is the case that a screener will notice the answer to a difficult answer, for example the study's sample is the wrong age, as soon as she begins reading the abstract. We do not suggest that the screener should "go back" to the previous items and mentally mark them as “yes” simply to complete the process. Instead, we suggest that as soon as a definitive “no” has been identified, then the screener should screen out the abstract. The process moves quickest when the easier items are identified first, but sometimes it does not work that way in practice.

3. Conduct introductory abstract screening trainings where screeners learn and pilot test the tool by screening the same 20 to 30 abstracts. Repeat as necessary until team reaches consensus.

After the abstract screening tool has been created, it will be distributed to the abstract screening team. The members of this team may or may not have experience screening abstracts. Regardless of the team members' experiences, however, it is critical to provide abstract screening training.

To conduct the training, the screeners should be familiar with the tool, the contents of the questions, and, preferably, why the questions are being asked. The leader of the training, the review manager, should describe thoroughly each question asked; no question should be overlooked or considered obvious. Even questions about the date of publication could be misinterpreted. For example, does the question “Was the study published after 1991?” include studies that were published in 1991 or only studies published in 1992 or later? Discussing each question with the abstract screening team prevents ambiguity which results in more accurate screening.

After a thorough discussion of the screening tool, the screeners should independently screen the same subset of abstracts using the screening tool.[20] The Cochrane Handbook[8] suggests pilot screening 10 to 12 abstracts in a training phase. The Agency for Health Research and Quality guidance[11] includes pilot screening of 10% to 20% of studies.

In our experience, 20 to 30 abstracts provided a sufficient number so that all screeners applied the inclusion criteria consistently. Frampton and colleagues also suggest including studies that are definitely eligible, unsure, and definitely ineligible in the pilot screening.

After each individual has screened the pilot abstracts, review team leaders should analyze the discrepancies. No matter how exact and unambiguous the screening tool, the screeners will disagree on which abstracts should be screened in or out. These disagreements, however, provide valuable information to the review team about limitations of the abstract screening tool because (a) the disagreements may point to poorly written questions and/or (b) (a) may provide valuable teaching opportunities. The disagreements should be discussed thoroughly prior to conducting further screening. The discussion may lead to a second round of piloting—this will depend on the experience of the screeners and complexity of the abstracts. The pilot phase of abstract screening should continue until every member of the team has had sufficient time to learn and understand the tool, the context, and the process of conducting abstract screening.

At the end of the pilot screening stage, the answers to the 20 to 30 pilot abstracts should be made widely available to screeners. The screeners can then refer to these answers during the screening process should questions arise. This can also be made available should additional screeners join the team and need to participate in the pilot training.

5 | DURING ABSTRACT SCREENING

4. Meet with the abstract screening team on a weekly or every other week basis.

After the initial training and piloting meetings end, and the full team begins abstract screening in earnest, the abstract screening team should meet on a weekly or every other week basis. The purpose of these meetings is to instill a culture of discussion, exploration, and curiosity while decreasing “coder drift.” As Lipsey and Wilson[13] suggested, coder drift occurs when screeners make individual decisions that differ from the group's decision-making process. Small decisions made individually over time, without correction, can result in unreliable decision-making by each screener. Meeting on a weekly or every other week basis reduces the risk of inaccurate individual decisions. To instill a culture of curiosity and understanding, the review managers should encourage questions and participation. It is sometimes difficult, however, to prompt questions from the screeners during the meeting. One practical option, therefore, is to ask screeners to write one specific question about a difficult abstract during the week and email it to the review manager. The review managers can then choose which questions need discussing with the group. The act of discussing one question can help spur other members of the team to ask questions or become curious about the abstracts they are screening.

Finally, regular meetings promote a sense of community among the screeners. Abstract screening, especially for more than 8 to 10 hours per week, can be isolating and tedious. Meeting with other individuals who are participating in the same work decreases feelings of isolation, increases buy-in, and ultimately promotes more efficient and effective screening.

5. Minimize changes to the screening tool.

The abstract screening tool, as explained above, should be piloted and modified at the beginning of the abstract screening process. As more individuals screen abstracts and work through the pilot round, clarifications to the abstract screening tool should be considered essential and beneficiary. Abstract screeners should feel empowered to suggest changes and ask for clarity.

Even the most rigorous pilot testing process, however, may result in questions that continue to lack clarity. That is why it is critical to meet with the abstract screening team weekly or at least every other week to discuss progress and any potential problems. During these meetings, it is tempting to make changes to the screening tool that impact ongoing, active screening. We suggest, however, that these changes be kept to an absolute minimum. We make this suggestion for several reasons. First, changing the screening tool in a substantive way naturally creates differences within the already screened studies. Although not vital to the end product, these changes impact what is reported in the PRISMA flow diagram.[22] Additionally, the changes may result in a less effective text-mining algorithm (see guideline 6). Second, changes to the tool can incite confusion, which in turn creates unreliable screening. Third, should changes become the norm instead of the exception, screeners may start to misunderstand what types of abstracts should be in or out. This may also decrease buy-in and participation, ultimately decreasing efficiency and effectiveness.

While we do not advocate for changes to the screening tool items, we do advocate for the inclusion of practical examples. As illustrated in our example tool in Online Appendix A, the screening items that address the abstract each specify examples to help guide the screeners' decision-making. Question 6 for example asks if the study uses a longitudinal design. Answering this question might be quite easy for more experienced reviewers, while less experienced reviewers may not know or recognize the various terminology used by study authors to indicate a longitudinal design. To clarify this question, we included various terms that might be used by study authors to represent a longitudinal design without using the phrase "longitudinal." In our example, we included the following words: "prospective, over time, trajectory, panel, waves, multiple time points, time 1, time 2, T1, T2, school transition." Providing this level of detail will ensure that the screeners understand the question and make reliable decision-making, while decreasing the probability of changing the screening item.

6. Use a text-mining abstract screening application.

Traditional abstract screening uses reference management software (such as EndNote or Zotero) or simple spreadsheets to list all citations for screening. The abstracts are then screened in the order that they were downloaded from the database searches. The first abstract screened is as equally likely to be kept for full-text screening as the last abstract. Structuring the process in this way can lead to boredom and unreliability as abstract screeners become increasingly comfortable with the material and less excited about the task at hand.

Using a text-mining abstract screening tool has the potential to eliminate or at least mitigate some of the problems associated with the traditional abstract screening process. Text-mining is a type of programming that provides computers with the ability to parse textual information without being explicitly programmed. A text-mining abstract screening application analyzes each abstract's textual information and the screening decision made by the screeners to understand the differences in textual information within a screened "in" and screened "out" abstract. The program then analyzes each additional abstract yet to be screened, posits the probability that each remaining abstract is eligible for the review based on the similarity to other abstracts previously screened, and then sorts the remaining abstracts by that probability of inclusion. The result is an ordered list of abstracts, where the abstracts with the highest probability of inclusion are at the beginning of the list, and the abstracts with the lowest probability of being included are at the end of the list. As a result, abstract screeners may move efficiently through the list of abstracts to screen as they move forward through the list because those studies most likely to meet the inclusion criteria are at the beginning. The increased efficiency encourages and motivates abstract screeners.

Text-mining abstract screening applications are readily available. Examples include Abstrackr,[23] Rayyan,[24] Covidance,[25] and EPPI Reviewer.[26] Other researchers have evaluated these programs' specificity and sensitivity.[27] Readers are encouraged to examine those articles to understand how and how well the various programs function. Olorisade, de Quincey, Brereton, and Andras[28] reported on a study attempting to compare the performance of different machine learning programs for citation screening but conclude that insufficient information is available to provide direct comparisons of their effectiveness.

This paper, therefore, seeks not to endorse one particular program. We recognize that some reviewers may prefer to conduct a systematic review in a contained database system. If this option is preferable, we recommend that reviewers use a program like EPPI Reviewer. However, if this is not a desired outcome, then one may use Abstrackr; we illustrate its functionality here because it is free, provides simple out-of-the-box functionality, and does not require learning an entire database system.

To begin, Abstrackr[29] allows users to create "projects" that warehouse all available citations. Once created, the user may upload a text file or Medline formatted reference document that delineates each citation's title, abstract, date of publication, and any other relevant information. When creating the project, users are asked a series of questions about the logistics of the project. The first question asks about the screening mode: single or double. We strongly suggest using double-screening for the purposes of training, accuracy, and reconciliation. We explain double-screening in more detail in guidelines 7 and 8. The second question asks about the order of the abstracts: random or most likely to be relevant. We suggest using the "most likely to be relevant" because this uses the text-mining functionality of sorting the citations by their probability of inclusion. The third question asks about the "pilot round size." The pilot round is when all screeners screen the same abstracts (see guideline 3). This allows for an easy analysis of discrepancies. The last question asks about "tag visibility," but it is not currently functional, and we will not discuss its use.

Using a text-mining tool such as Abstrackr also allows for easy project management. Under the "Admin" tab, the review manager can add or remove screeners as well as give administrative privileges to other participants. That same tab allows managers to assign abstracts to screeners. The "upload terms" tab, under the Admin tab, allows a manager to input certain terms that will be highlighted in the abstract when users are screening. For example, if the words "longitudinal" or "bullying" are important, adding that "term" to the list will highlight it in green text. Screeners can also add antithesis words, for example the words "cross-sectional" or "qualitative," and they will appear highlighted in red. Adding many of these terms provides further reassurances and ease of use to the screeners, and they improve the text-mining functionality.

From the screener's point of view, all project work is contained and easy to access. Clicking on the "screen" button takes screeners to the next available abstract to screen. Once there, the screener uses the abstract screening tool (see guidelines 1 and 2) to answer questions about the title and abstract. The green "check mark" indicates that an abstract should be included; the red "x" indicates that an abstract should be dropped. The "review labels" button allows the screener to view all decisions made and make changes to those decisions, if needed. The counter at the bottom indicates how many abstracts have been screened. Especially important, the program is browser-based and can be accessed via a computer, tablet, or smartphone. We support the screening of abstracts, once the screening tool has been sufficiently memorized, in any location where the screener feels comfortable.

Once screening begins in earnest, the review manager may observe how many abstracts that Abstrackr "predicts" will be included in the remaining abstracts to screen. The "predictions" button, located in the "My projects" page, renders a histogram of all remaining probabilities; the number Abstrackr predicts will be included is the number of abstracts that have a probability of inclusion greater than 50%. We do *not* suggest stopping abstract screening once the counter reaches "0" because this would mean that Abstrackr has perfect prediction. Although Rathbone, Hoffman, and Glasziou's (2015) research indicated that *Abstrackr* has a high accuracy rate, limited work has been conducted on this topic to date. Moreover, should changes to the screening tool be made during the screening process, this prediction function may not represent an accurate count. Therefore, until more evidence is available, we suggest screening all available abstracts.

7. Require independent double-screening of each abstract.

Double-screening all available abstracts is not a new practice and has been suggested as a best practice for decades.[18] Guidance from the Cochrane Collaboration,[8] the Institute of Medicine,[10] and the Center for Review and Dissemination[9] all include the importance of independent screening of identified studies by at least two independent coders. Single-screening has the potential to remove studies from consideration before they can be vetted fully. It is simply too easy to make a mistake and remove a study.

Simply implementing double-screening without careful oversight, however, falls short of the managerial requirements that large-evidence reviews need. We suggest that review managers use the data generated from double-screening to guide future decisions and training. In Abstrackr, for example, a manager may download a spreadsheet that delineates every abstract screened as well as the screeners who made decisions and their choices. A review manager, therefore, may use this information to calculate agreement rates as a group or by the individual. Individuals who have high levels of disagreement may require booster training. High group disagreement rates, say less than 75% agreement, on the other hand, may be indicative of a systemic problem with the screening tool or the training. Regardless of the reason for disagreement, our suggestion is that screening should be monitored continuously.

One other suggestion is that the results of the abstract screening process should regularly be made available to the abstract screening team. We do not suggest publishing individuals' disagreement rates. Instead, sharing the raw sum or percentage of total abstracts screened may incentivize screeners to continue. We recommend quantifying and sending information out to screeners at least once per week; review managers leading many screeners (eg, more than three to five individuals) should consider sending updates out more regularly.

8. Reconcile disagreements throughout the abstract screening process.

No matter how effective the screening tool is, or how often the abstract screening team meets, screening disagreements will occur. Sometimes, these are due to simple human error; other times, they are due to “coder drift” or more systemic issues with the interpretation of the screening tool questions. Difficult abstracts where information is lacking prevent perfect agreement as well. We have found that it is common practice to reconcile these disagreements after all abstracts have been screened. For small projects, for example 300 to 500 abstracts, this is a fine practice. For larger projects, however, we suggest reconciliation occur after only 20% to 30% of the abstracts have been screened. More frequent reconciliation limits the need to re-screen abstracts due to potential errors. Afterwards, reconciliation should continue after screeners complete each additional 20% to 30%.

We make this recommendation because reconciliation can be influential in ensuring that abstract screeners make the correct determination throughout the screening process. Reconciliation forces screeners to read another screener's decision that is different from their own, decide whether they support the decision, defend their own decision, and then make a final determination. All of these processes force the screener to think carefully about the abstract as well as the screening questions.

We also suggest that reconciliation occur relatively early on in the abstract screening process because we assume that the screening team uses a text-mining screening tool that sorts the abstracts according to their relevance. The result is that the more difficult abstracts, and the ones that are likely to be included in the review, tend to be listed toward the front of the screening process. Waiting until the end of the screening process decreases the impact on screening process because the abstracts at the end of the process are, by nature, less likely to be relevant to the review. Difficult abstracts, where the two screeners cannot come to a consensus, should be discussed with a third screener. The third screener is most often the review manager or team leader. It is often helpful to discuss difficult abstracts with the group during team meetings, especially if a decision cannot be made by the original three screeners.

In our review project, 15 screeners had screened 1213 abstracts by the end of the first week of screening, with 3588 abstracts screened by the end of the second week. At this point, the leadership team noticed a potential issue while tracking disagreements. At the end of the third week, the screening team had screened 6249 abstracts but disagreed 30% of the time. After a day of consultation, the decision was made to halt screening and conduct an initial reconciliation. The abstract screening team met with the review leadership to discuss the reconciliation process and make any necessary changes to the abstract screening tool. The project director sent each screener a spreadsheet that listed every abstract where they disagreed with another screener. Independently, the two abstract screeners reviewed each decision and then contacted the other screener to determine a final rating. Some abstracts required a third screener to review the abstract. The most difficult abstracts were discussed with the group.

After 1 week of conducting reconciliation among pairs of screeners, the disagreement rate decreased to 15.2%. Review leadership determined that any person who completed reconciliation could continue with abstract screening. After reconciliation, the disagreement rate continued to decrease while the number of abstracts screened increased. At the end of the screening process, the final disagreement rate was 8.2%.

Finally, we recognize that reconciling throughout the abstract screening process will render traditional reliability statistics insensible. Many systematic reviews provide a percentage of times that coders agreed when making screening decisions, and these traditional reliability statistics require that the entire abstract screening process be complete prior to their calculation. Should reconciliation occur throughout the abstract screening process, calculation of these statistics will not be possible. We argue, however, that the gain in efficiency and reliability outweighs the potential consequences of not being able to report these traditional statistics.

9. Encourage screeners by limiting time on task, promoting intellectual buy-in, and providing incentives.

As we have noted, abstract screening is an arduous and thankless task. Therefore, review managers, not unlike

TABLE 3 Time estimation based on handling of reviewer discrepancies

Review Size	Number in Dispute	Resolution Time	Number to Retrieve	Retrieval Time	Dispute + Retrieval	Screening Time	Total Time
Retrieve all disputes							
100	10	0	10	1.25	1.25	3.33	4.58
1000	100	0	100	12.50	12.50	33.33	45.83
2500	250	0	250	31.25	31.25	83.33	114.58
5000	500	0	500	62.50	62.50	166.67	229.17
7500	750	0	750	93.75	93.75	250.00	343.75
10 000	1000	0	1000	125.00	125.00	333.33	458.33
Resolve all disputes							
100	10	1.67	5	0.63	2.29	1.67	3.96
1000	100	16.67	50	6.25	22.92	16.67	39.58
2500	250	41.67	125	15.63	57.29	41.67	98.96
5000	500	83.33	250	31.25	114.58	83.33	197.92
7500	750	125	375	46.88	171.88	125.00	296.88
10 000	1000	166.67	500	62.50	229.17	166.67	395.83

Notes: We assume 10 min per dispute resolution; 7.5 min per article retrieval; 20 min per article screening. All time amounts represented in hours.

managers in other workforce sectors, must work tirelessly to motivate abstract screeners to continue to screen on time and efficiently. Also similar to other workforce sectors is the means to motivate abstract screeners. We have found that two techniques produce particularly effective results.*

The first technique is simply to limit screeners daily time on task. This has been referred to in the literature as conducting a task by doing so in "bursts."[30] The theory is that the screener should only work for short periods of time to maximize engagement and concentration. Attempting to screen for hours at a time, for example more than 3 hours per day, may lead to unreliability, slowness, or simple burnout. In our reviews, we strongly suggest to screeners that the maximum amount of time they should screen at any one time is 2 hours per session. Because we involve researchers who need to bill hourly, screeners do sometimes screen more than 3 hours per day. But as a rule, we attempt to limit these sessions as much as possible.

The second technique is to encourage intellectual buy-in. This is especially relevant if the screening team consists of university students or individuals who represent a particular interest in the topic area. We have found it particularly effective to encourage screeners to consider ways to use the database for tangential projects. For example, a review on the effects of middle school math interventions may be used to start a project on the effects of elementary school math interventions. An additional technique is to include the screeners in the decision-making process during the creation of the abstract screening tool and ensure that the screeners' concerns are heard by the review managers. All of these techniques can be used to promote buy-in.

A third technique is to use incentives. As economists point out, an economic incentive is one that has the potential to change behavior.[31] Should the project have the resources, several options are available. A simple option is to create a small contest that awards prizes to screeners who (a) screen the most abstracts, (b) have the high agreement rate, or (c) log the most screening time. Again, we lack empirical evidence to support our claims, but anecdotally, even small economic incentives drive productivity. They also instill a team-building and collegial atmosphere that promotes discussion and participation. A good review manager seeks to engender these behaviors.

6 | AFTER SCREENING ENDS

10. Analyze the process and decisions after screening has been completed.

The end result of the abstract screening process is a spreadsheet that includes decisions for every citation found. Completing abstract screening, especially for large-evidence projects, has the potential to feel like a major accomplishment.

*Note: we lack empirical evidence of their effectiveness. The use of the word "effective" is in the anecdotal sense. We also believe that these incentives might be effective for paid staff as well as unpaid research assistants.

Review managers, however, also tend to lose sight of the process and decisions made by the abstract screening team because the next steps in the systematic review process await. If the ongoing project is the only one planned, then it may be reasonable to move on to the next step. Should additional projects be in the works, however, it is important to analyze the screening decisions and determine what parts of the process worked.

Conducting a postmortem of the abstract screening process is akin to debriefing a research participant or analyzing exit poll results. The purpose is to understand what worked, what did not work, and how the process could be improved in the future. For example, some abstract screening tools track when abstract screening decisions were made. One way to analyze the results is to observe whether abstract screeners agreed more (or less) over time. Abstract screening that improves over time indicates that the process worked; more disagreements over time might indicate that greater emphasis be placed on "coder drift" and possibly more reconciliation stoppages. Analyzing the results, and sharing those results with the review team, also ensures that abstract screening records are maintained and available, should they be needed in the future. Once the review has been completed, the abstract screening records will need to be reported. Analyzing the results ensures the records are orderly and available.

7 | CONCLUSION

The purpose of this paper was to provide review teams and managers of large-evidence reviews with a set of practical abstract screening guidelines. Our guidelines help to ensure that the abstract screening process concludes swiftly and with as few errors as possible. Future research is still required to evaluate some of our claims, yet we believe that these guidelines should be made available to the research community at-large and their use will promote effective research syntheses.

Some of our suggestions, it should be noted, will have greater impact as the review size increases. We suggest in guideline 8, for example, that review authors should direct their team to reconcile disagreements throughout the screening process. This may seem counterintuitive and some might suggest retrieving all articles in disagreement rather than resolving disagreements before deciding which articles need retrieval. To illustrate how this decision impacts the amount of time required, we conducted a brief time analysis delineated in Table 3. The rows in the top portion, labeled "Retrieve All Disputes," assume that the review team will retrieve every article in dispute. The rows in the lower portion assume that the review team will resolve discrepancies to determine the articles needed for retrieval. We also assume that (a) for each abstract in dispute, it will take the review team 10 minutes to resolve the dispute; (b) each article will take approximately 7.5 minutes to retrieve; and (c) each article will require 20 minutes to full-text screen.

The final column in the table illustrates the total amount of time required, from dispute resolution through full-text screening, for the various review sizes. For a review with 100 to 1000 citations found, the difference in total time is minimal and perhaps within a natural variation range (1-5 hours total time difference). As the size of the review increases, however, the difference in the total time required begins to increase. For the largest-scale review ($n = 10\,000$ citations found), the difference in total time is 62.5 hours. This represents a difference of nearly $1700.00 USD if we assume a review team member's hourly rate is around $25.00 USD per hour. Clearly, seemingly small decisions made at scale can have a lasting impact.

The large-evidence abstract screening process is a tedious and thankless task. It requires multiple individuals, knowledgeable about a particular topic, comb through an endless list of abstracts that may or (likely) may not fit the inclusion criteria. Review managers of staffs larger than three or four screeners must stay abreast of the progress, ensuring that drift is minimized, sufficient agreement remains, and motivation is maintained. Through dedicated processes and consistent oversight, the review manager can safeguard against inefficiencies and inaccurate decision-making. As extant literatures continue to grow, these best practice guidelines should prove helpful to researchers and review managers in the future.

ACKNOWLEDGEMENTS

The authors wish to thank Jeff Valentine, Sandra Jo Wilson, and Mark Lipsey for inspiring this work and providing years of informal guidance.

DATA AVAILABILITY STATEMENT

Data supporting this publication is available upon request from the first author.

Keywords

abstract screening, best practices, information retrieval, meta-analysis

ORCID

Joshua R. Polanin https://orcid.org/0000-0001-5100-0164

Joshua R. Polanin[1]
Terri D. Pigott[2]
Dorothy L. Espelage[3]
Jennifer K. Grotpeter[4]
[1]American Institutes for Research, Washington, DC, USA
[2]Loyola University Chicago, Chicago, Illinois, USA
[3]University of Florida, Gainesville, Florida, USA
[4]Development Services Group, Inc, Bethesda, Maryland, USA

Correspondence
Joshua R. Polanin, American Institutes for Research, Washington, DC.
Email: jpolanin@air.org

REFERENCES

1. Pigott TD. Advances in meta-analysis. *Stat Soc Behav Sci*. 2012; xiii:155. https://doi.org/10.1007/978-1-4614-2278-5
2. Wrzus C, Hänel M, Wagner J, Neyer FJ. Social network changes and life events across the life span: a meta-analysis. *Psychol Bull*. 2013;139(1):53-80. https://doi.org/10.1037/a0028601
3. Kupers E, Lehmann-Wermser A, McPherson G, van Geert P. Children's creativity: a theoretical framework and systematic review. *Rev Educ Res*. 2019;89(1):93-124. https://doi.org/10.3102/0034654318815707
4. Wilson DB, Brennan I, Olaghere AA. Campbell systematic review 2018:5 crime and justice coordinating group police-initiated diversion for youth to prevent future delinquent behavior: a systematic review. *Campbell Syst Rev*. 2018;5:1-85. (June). https://doi.org/10.4073/csr.2018.5
5. Birks JS, Harvey RJ. Donepezil for dementia due to Alzheimer's disease. *Cochrane Database Syst Rev*. June 2018. https://doi.org/10.1002/14651858.CD001190.pub3
6. Reed JG, Baxter PM. Using reference databases. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. New York, NY: Russell Sage Foundation; 2009:73-93.
7. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545. https://doi.org/10.1136/bmjopen-2016-012545
8. Higgins J, Deeks J. Chapter 7: Selecting studies and collecting data. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). www.handbook.cochrane.org. Published 2011.
9. Tacconelli E. Systematic reviews: CRD's guidance for undertaking reviews in health care. *Lancet Infect Dis*. 2010;10(4):226. https://doi.org/10.1016/S1473-3099(10)70065-7
10. Medicine I of. *Find What Works In Health Care. Standards for Systematic Reviews*.; 2011. https://www.nihlibrary.nih.gov/sites/default/files/Finding_What_Works_in_Health_Care_Standards_for_Systematic_Reviews_IOM_2011.pdf.
11. McDonagh M, Peterson K, Raina P, Chang S, Shekelle P. Methods guide for effectiveness and comparative effectiveness reviews.; 2013. https://www.ncbi.nlm.nih.gov/books/NBK126701/?report=classic.
12. Cooper H. *Research Synthesis and Meta-Analysis: A Step-By-Step Approach*. Thousand Oaks, CA: Sage Publ; 2009:360.
13. Lipsey MW, Wilson DB. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications, Inc; 2001 http://psycnet.apa.org/psycinfo/2000-16602-000. Accessed May 16, 2017.
14. Olofsson H, Brolund A, Hellberg C, et al. Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Res Synth Methods*. 2017;8(3):275-280. https://doi.org/10.1002/jrsm.1237
15. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Syst Rev*. 2015; 4(1):80. https://doi.org/10.1186/s13643-015-0067-6
16. Shemilt I, Simon A, Hollands GJ, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Res Synth Methods*. 2014;5(1):31-49. https://doi.org/10.1002/jrsm.1093
17. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*. 2010;7(9):e1000326. https://doi.org/10.1371/journal.pmed.1000326
18. Rosenthal R. *Meta-Analytic Procedures for Social Research*. 2nd ed. Newbury Park, CA: Sage Publications, Inc; 1991.
19. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*. 2015;350(jan02 1): g7647. https://doi.org/10.1136/BMJ.G7647.
20. Livoreil B, Glanville J, Haddaway NR, et al. Systematic searching for environmental evidence using multiple tools and sources. *Environ Evid*. 2017;6(1):23. https://doi.org/10.1186/s13750-017-0099-6
21. Brunton V, Stansfield C, Caird J, Thomas J. Finding relevant studies. *Gough, D Oliver, S Thomas, J, An Introd to Syst Rev* Sage London. April 2017. http://discovery.ucl.ac.uk/1542859/. Accessed February 11, 2019.
22. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol*. 2009;62(10):e1-e34. https://doi.org/10.1016/j.jclinepi.2009.06.006.
23. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center. In: *Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics—IHI'12*. Vol.819 New York, New York, USA: ACM Press; 2012 https://doi.org/10.1145/2110363.2110464.
24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016; 5(1):210. https://doi.org/10.1186/s13643-016-0384-4
25. Covidance. Covidence—accelerate your systematic review. https://www.covidence.org/about-us. Published 2017. Accessed November 28, 2017.
26. Thomas J, Brunton J, Graziosi S. *EPPI-Reviewer 4: Software for Research Synthesis*. London, England: Social Science Research Unit, UCL Institute of Education; 2010.

27. Saha TK, Ouzzani M, Hammady HM, Elmagarmid AK. A large scale study of SVM based methods for abstract screening in systematic reviews. In: *Document Analysis and Recognition (ICDAR)*. Washington, DC: ArXiv preprint; 2013 http://arxiv.org/abs/1610.00192. Accessed November 28, 2017.

28. Olorisade BK, de Quincey E, Brereton P, Andras P. A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. In: *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering—EASE'16*. ; 2016:1–11. https://doi.org/10.1145/2915970.2915982.

29. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol*. 2017;91:31-37. https://doi.org/10.1016/j.jclinepi.2017.08.011

30. Croft A, Vassallo D, Army MR-J of the R, 1999 undefined. Handsearching the journal of the royal army medical corps for trials. *researchgate.net*. https://www.researchgate.net/profile/Ashley_Croft/publication/12879894_Handsearching_the_Journal_of_the_Royal_Army_Medical_Corps_for_Trials/links/0f31752dff9c88a8e7000000/Handsearching-the-Journal-of-the-Royal-Army-Medical-Corps-for-Trials.pdf. Accessed March 30, 2019.

31. Gneezy U, Meier S, Rey-Biel P. When and why incentives (don't) work to modify behavior. *J Econ Perspect*. 2011;25(4):191-210. https://doi.org/10.1257/jep.25.4.191

SUPPORTING INFORMATION

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