Methods in Ecology and Evolution

DR YONG ZHI FOO (Orcid ID : 0000-0001-7627-2991)
DR ROSE ELEANOR O'DEA (Orcid ID : 0000-0001-8177-5075)
DR JULIA KORICHEVA (Orcid ID : 0000-0002-9033-0171)
DR SHINICHI NAKAGAWA (Orcid ID : 0000-0002-7765-5182)

Article type : Review
Editor : Thomas White

A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution

Yong Zhi Foo[1], Rose E. O'Dea[1], Julia Koricheva[2], Shinichi Nakagawa[1*], Malgorzata Lagisz[1*]

[1]Evolution and Ecology Research Centre, School of Biological and Environmental Sciences, University of New South Wales, Sydney, 2052, NSW, Australia
[2]Department of Biological Sciences, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK

*These authors contributed equally as the last authors.

Abstract

1. Well-conducted systematic reviews are invaluable for synthesizing research findings. The conclusions of a review depend on how the research question was formulated, how relevant studies were found, and how studies were selected for synthesis.

2. Here, we present a practical guide for ecologists and evolutionary biologists on formulating a question for a systematic review, and finding a representative sample of research findings.

3. We explain the steps involved using a worked example and practical training exercises. Throughout this guide we share tricks of the trade, included rules of thumb and software that we have found useful.

4. We hope our paper helps demystify the systematic search process and encourages more researchers to adopt a systematic and reproducible approach when searching the literature.

keywords: *systematic review, meta-analysis, narrative review, systematic search, screening, Boolean*

Introduction

The goal of a systematic review is to provide valid summary of primary research findings through a pre-planned and explicit procedure (Moher, Stewart, et al., 2015). Research findings can be summarised in different ways, including meta-analyses, narrative reviews, or evidence maps. Regardless of the methods of synthesis, all systematic reviews need to find an unbiased sample of available evidence to accurately reflect the state of our knowledge on a topic. Therefore, all systematic reviews follow planned and logical steps to gather and select research findings (Fig. 1).

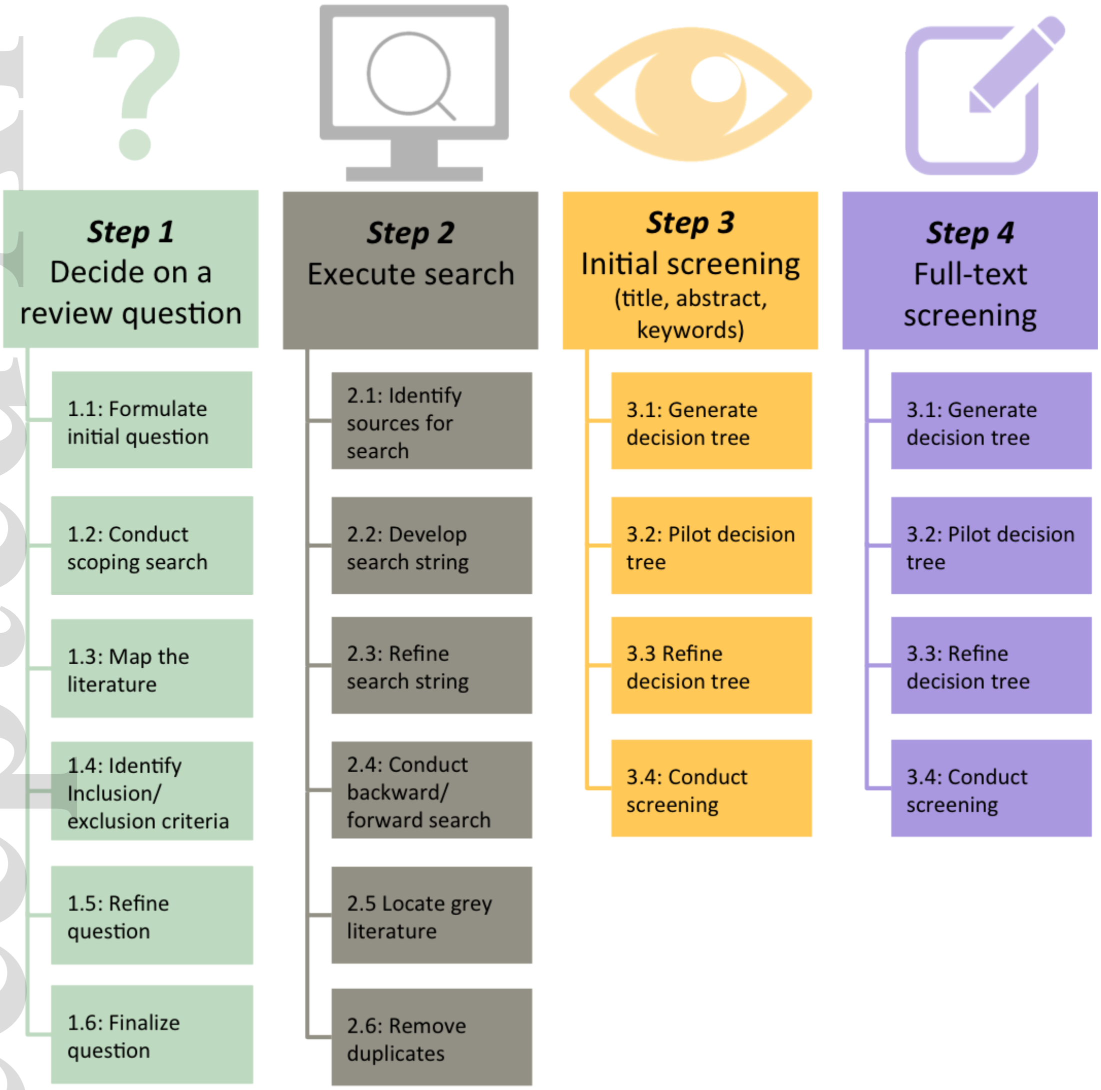

Figure 1. Summary of the typical steps in a systematic search.

Practical pitfalls of a systematic search

Although the steps of a systematic search (Fig. 1) may seem straightforward, they are often difficult to execute for two broad reasons. First, there are many decisions to make that could affect the outcome of the review (e.g., how broad should the review question and inclusion criteria be, and which literature sources to use?). Second, systematic review methods are iterative (Booth et al., 2012). Rather than steps proceeding linearly, previous steps are often returned to and revised (e.g., the review question might need to be narrowed after an initial search returns too many results). Before committing to a particular review question or search method, therefore, we recommend extensively piloting methods for searching and screening (Fig. 2).

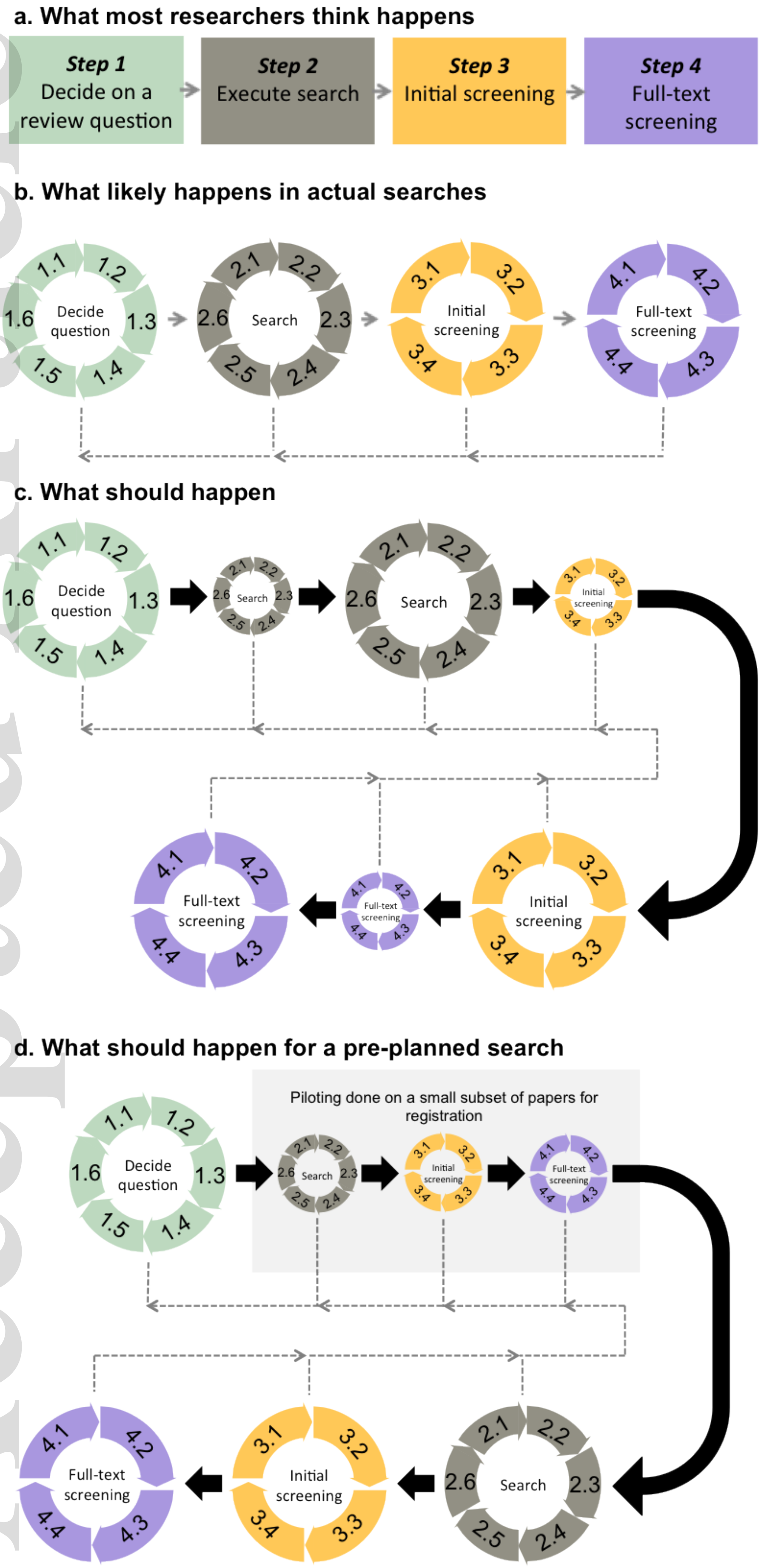
a. What most researchers think happens
Step 1
Decide on a review question
Step 2
Execute search
Step 3
Initial screening
Step 4
Full-text screening
b. What likely happens in actual searches
Decide question
Search
Initial screening
Full-text screening
c. What should happen
Decide question
Search
Initial screening
Full-text screening
d. What should happen for a pre-planned search
Piloting done on a small subset of papers for registration
Decide question
Search
Initial screening
Full-text screening

Figure 2. The search and screening process: What researchers want to happen (a), what likely happens instead (b), and how this could be avoided (c, d). Sub-steps are folded into a circle in (b, c, d) to illustrate that the sub-steps themselves can be iterative. The dotted arrows point backwards to illustrate revisiting earlier steps, if needed. Smaller circles represent pilot tests. Registered systematic reviews, especially registered meta-analyses, usually require pilot-testing before registration and the conduct of the actual search (d).

Aim and structure of this paper

We aim to provide a practical guide on conducting systematic searches for beginners, and include practical exercises in the Supplementary Materials. Because question formulation and study screening are integral to the methods of a systematic search (Fig. 2), we provide detailed guidance for each of these steps. Within each step, to demonstrate our guidance, we include a worked example from evolutionary biology (from a Stage-1-accepted registered meta-analysis on the "terminal investment hypothesis").

Deciding on a review question

A well-formulated question informs subsequent steps of the search process. The question should be sufficiently general to address the topic of interest, but not so broad that the search becomes impractical. Deciding on the appropriate question involves five sub-steps: 1) formulating the initial question, 2) conducting a scoping search, 3) mapping the literature, 4) identifying the inclusion criteria, 5) refining the question, and 6) finalizing the question.

Formulating the initial question

We might begin with a broad topic, before narrowing it down (e.g. to specific taxa or study designs). For example, trade-offs between reproduction and survival can be narrowed to paternal effort and survival in birds (Santos & Nakagawa, 2012) or comparing seed production between annual and perennial plants (Vico et al., 2016). Alternatively, if too few studies are available to address a specific research question, the scope might be expanded. Often, the question will undergo multiple rounds of broadening and narrowing. For review topics that have applied implications it is important to solicit the input of stakeholders at this stage (Haddaway et al., 2017).

Scoping search

The purpose of a scoping search is to familiarise oneself with a particular topic, determine whether there is sufficient primary research to conduct a review, and identify existing reviews (Khan et al., 2011; Lipsey & Wilson, 2001). Scoping searches gather a small set of the most relevant articles, and can help identify experts on the topic for consultation. Recent narrative reviews are very useful for estimating the amount of research available and identifying landmark empirical studies. The minimum number of studies needed depends on the scope of the question, types of systematic review, and practical considerations such as PhD deadlines.

Scoping searches can be done using online reference platforms like Web of Science or Scopus, or even search engines such as Google Scholar. Google Scholar orders search results based on their potential relevance to the search terms, which can be useful for locating articles quickly. But Google Scholar is subject to drawbacks that make it unsuitable as the main source of the actual literature search (Gusenbauer & Haddaway, 2020; see section below on Grey Literature).

Map out the literature

Literature mapping is a useful tool for using articles identified in scoping searches to understand the wider literature (Cobo et al., 2011; Zupic & Čater, 2015). For example, you could organize author keywords into categories or generate word clouds from the titles and abstracts to identify important terms. Doing so helps to reveal clusters of theoretical frameworks, typical research approaches, and commonly used populations (e.g., taxa, locations) that could be the focus of the review.

Identify the question

Once familiar with the literature, you can specify your research question using the PICO/PECO framework (Richardson et al., 1995), which identifies four components in a research question: **P**opulation, **I**ntervention/**E**xposure, **C**omparator, **O**utcome. The framework was formulated originally for clinical experiments. Therefore, not all of its components are applicable to all studies in ecology and evolution. For instance, in correlational studies, where researchers are interested in whether A is correlated with B, only the population and the outcome (whether a correlation or prevalence is assessed) are relevant (i.e. **PO**). If a meta-analysis is intended as part of the systematic review, the

PICO/PECO framework can help identify suitable effect sizes (e.g., Hedges' *g* and response ratio for PICO/PECO and Fisher's *Zr* for PO).

Refine the question, identify inclusion criteria and finalize question

Questions can be refined by adjusting individual PICO/PECO components. Refining a question might be necessary to fit the systematic review into a funding deadline, or for theoretical reasons. For instance, researchers might want to focus only on experimental studies to establish causal relationships (e.g., Roberts, Buchanan, & Evans, 2004). A clear question framework translates easily into inclusion/exclusion criteria for which studies will be included in the research synthesis. A question might require multiple rounds of refining before being finalized. Before we dive into finding and screening studies, let us introduce the worked example.

The terminal investment hypothesis: deciding on the review question

This worked example is based on a Stage-1-accepted registered report at the journal *BMC Biology* (therefore, in this worked example, "we" refers to the registered report authors: Y.Z.F., R.E.O., S.N., and M.L.). The terminal investment hypothesis states that when individuals perceive a sufficiently intense survival threat, they might increase current reproduction to make the most of their remaining reproductive potential (Clutton-Brock, 1984; Duffield et al., 2017; Fisher, 1930, see Fig. 3 for theoretical background).

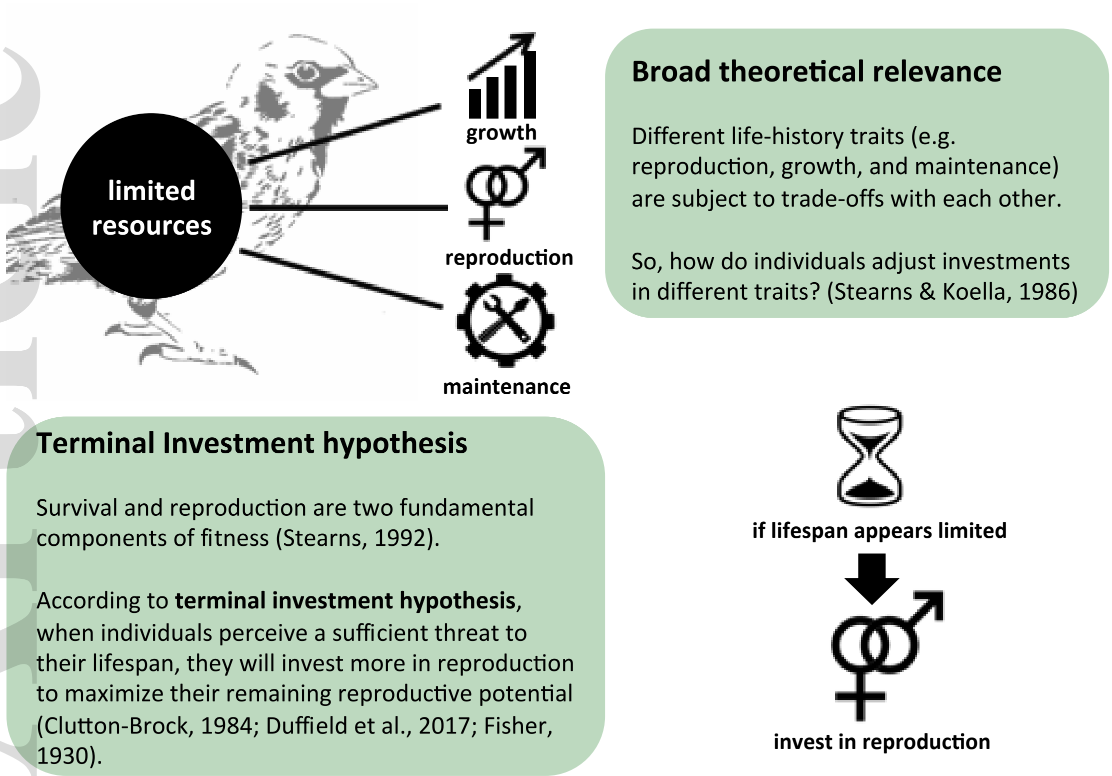

Figure 3. Theoretical background of the terminal investment hypothesis

Scoping search

The general question is whether individuals perceiving a survival threat invest more in reproduction. To refine our review question, we used a scoping search to understand how researchers typically investigate the terminal investment hypothesis empirically. We entered *"terminal investment" review* into Google Scholar (the quotation marks are used for phrases). We found a number of narrative reviews on the topic. The first five pages of the search results also revealed a substantial number of empirical studies testing the hypothesis.

Map out the literature

We took 16 of the empirical studies that we found (listed in Table S1) to map out the literature by organising author keywords into major categories (Fig. 4a). We also generated a word cloud from their titles and abstracts (Fig. 4b) using the R package *wordcloud* (Fellows, 2018). Fig. 4 shows that studies investigated numerous reproductive investment variables (e.g., egg size, male pre- and post-copulatory traits, parental care, mating effort, behavioural displays). Words such as "immune" and "infection" were common in titles and abstracts, pointing to a common experimental approach: using an immune challenge as a survival threat.

Our observations were corroborated by a recent review by Duffield and colleagues (2017), which identified over 50 primary research papers on the topic. The studies tested a wide range of reproductive investment variables. Most of them applied immune challenges as a survival threat. Importantly, we did not find any published quantitative reviews or meta-analyses, pointing to a gap in the literature.

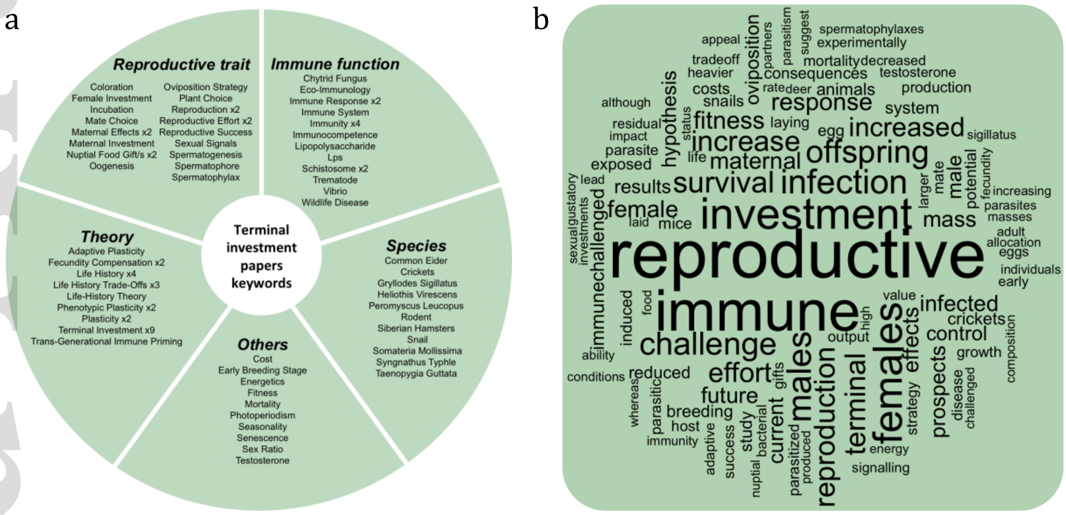

Figure 4. Scoping of the initial set of relevant papers on the terminal investment hypothesis: (a) article keywords organized into categories and (b) word cloud of titles and abstracts.

Identify question

The review by Duffield et al. (2017), together with the 16 empirical studies that we identified, assured us that there was sufficient primary research for a meta-analysis. Therefore, we decided to conduct a systematic review and meta-analysis on whether an immune survival challenge increases reproductive investment. Our question fit within the PICO framework, as follows:

__P__opulation: Any species

__I__ntervention: Immune challenge

__C__omparator: No challenge

__O__utcome: Reproductive investment measures

Refine question, identify exclusion/inclusion criteria, and finalize question

We used our scoping search results to refine our question. From Fig. 4, and the review by Duffield et al. (2017), we saw most studies were done on animal or insect species. Furthermore, plants have very different life histories and immune systems compared to the animals. So, we decided to limit our **P**opulation to multicellular non-plant species. We also saw that the immune challenges included both live pathogens and non-pathogenic (non-live) substrates. However, live pathogens can sometimes alter host life-history traits, including reproductive ones, which may confound our results. Therefore, we limited our **I**ntervention to non-pathogenic immune challenges. To be able to establish causal link, we decided to include only experimental studies that used an immune challenge as a manipulation, and not quasi-experiments comparing individuals that were already parasitized *vs*. those that were not. Our finalized PICO components were:

P*opulation: multicellular non-plant species*

I*ntervention: experimental immune challenge using non-pathogenic substrates*

C*omparison/Control group: unchallenged group of animals, otherwise in the same state*

O*utcome: any reproduction-related traits*

Try it yourself!

Try exercises 1 'Formulating the question for a meta-analysis' and 2 'Deciding on inclusion criteria' in the Supplementary Material.

Executing the search

Searches should capture as many relevant studies as possible while reducing the number of irrelevant ones. To achieve this goal, there are six sub-steps to consider: 1) identifying the most appropriate literature sources, 2) generating the search string for database/platform searches, 3) refining the search string, 4) supplementing the database search by examining the reference lists and citing articles of relevant studies and reviews (also known as backward and forward searches), 5) searching for grey literature, and 6) removing duplicates.

Identify literature sources

There are many search methods to identify relevant studies, each with their own strengths and weaknesses, including: database searches; web searches; citation-based searches; and expert-based searches (including asking experts, and searching publication histories). At present, online databases of published papers provide the greatest coverage of research findings and, therefore, represent the primary resource in most systematic reviews. Web of Science and Scopus are commonly used in ecology and evolution, although other databases might be more relevant depending on the topic (see Gusenbauer & Haddaway, 2020 for list of potential databases), and not all researchers have institutional access to these tools. Note that some "databases" are actually platforms for access to a collection of databases. For example, Web of Science is a platform providing access to Web of Science Core Collection and over 10 other databases. Different institutions may have subscriptions to only a subset of databases. Therefore, when using such platforms, you should check or consult with a librarian in order to select the relevant databases, which can help reduce the number of irrelevant results.

You should gather publications from multiple (at least two) databases because overlaps between published databases can range from around 50% to as low as 11% (Bar-Ilan, 2018; Mongeon & Paul-Hus, 2016). For researchers who have limited access to the required databases, they might need to collaborate with others who do have access or supplement their search using other search methods.

Develop search string for database/platform searches

To generate an effective search string, researchers need to understand the "language" of Boolean and other search operators (e.g., wildcards). Boolean search combines multiple search terms and phrases (Fig. 5; Hjørland, 2014), thus increasing the efficiency and reproducibility of searches. Search results are based on the contents of the database records. If search terms are being used in different contexts, the records found may not necessarily be relevant to the topic of interest. For example, searching for "sexual selection" on Web of Science reveals results from not just evolutionary-biology-related fields, but also others like education and music.

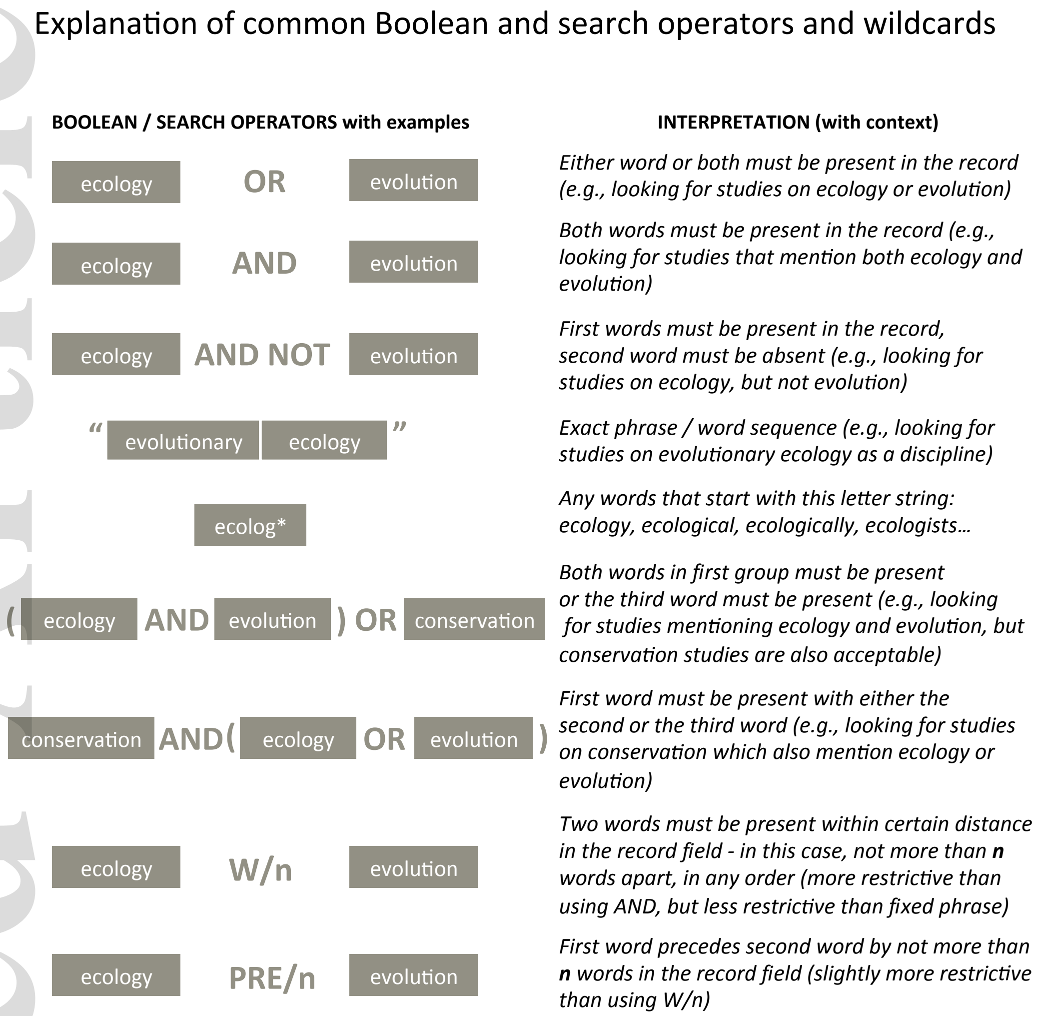

Figure 5. Basic Boolean operators, wildcards, and phrases used in database searches (Scopus convention). Realistic examples of their use "in action" is presented in Fig. 6. Note that, in advanced searches, round brackets allow easy and clear hierarchical nesting of multiple search elements. W/n and PRE/n are called "proximity operators". A note of caution: when using the AND NOT operator, check the list of excluded papers to ensure that you do not accidentally exclude relevant papers due to double negatives (e.g., using "AND NOT vertebrates" in order to focus on invertebrates might lead us to exclude papers that mentioned "non vertebrates").

An effective search, therefore, involves identifying words and phrases, and their synonyms, that are present in database records of relevant papers ("inclusion terms"), or are present in irrelevant records that happened to be captured by our inclusion terms ("exclusion terms"). Inclusion terms are usually linked to theory, or one of the PICO components (the target population, the manipulation, the outcome, etc.). You can also identify them from the literature mapping results (Fig. 4) and by using topic-modelling tools such as the R packages *scimeetr* (Rivest, 2016) and *revtools* (Westgate, 2019). Emerging developments in automated search string generators look very promising (e.g., Grames, Stillman, Tingley, & Elphick, 2019). However, such technology is still in the early stages of development and more data is required to determine the wider applicability to ecology and evolution, where terminology may be inconsistent across taxa and experimental designs. Thus, we caution against blind reliance on these tools.

Refine search string

Search strings are refined through pilot screening, and measuring the percentage of relevant records (i.e. the hit rate or precision), the percentage of known relevant records that are not found in the search (i.e. the miss rate), and the total number of records retrieved. We recommend running a pilot screening on a random sub-sample of resulting records to estimate the hit rate. Based on our experience, a random selection of 100 records is a reasonable number for this purpose. We try to aim for a hit rate of at least 10%, but this is not always possible (e.g., for narrow inclusion criteria). The miss rate should be minimised, and can be measured by comparing the search record against a core set of relevant papers (e.g., from your scoping search or from a previous review) to see if the search captures the majority of the set. The search can be iteratively refined by adding inclusion terms from missed studies, or exclusion terms from irrelevant studies.

The recommended total number of records retrieved depends on a number of factors, including the topic of review, time available, and amount of help from collaborators and research assistants. Between 1000 and 3000 records might be a reasonable starting target for many questions, as long as pilot searches reasonably indicate that most of the relevant records can be retrieved. Our first author (Y.Z.F.) once screened through > 100,000 abstracts and titles to find only 122 papers for one of his meta-analyses (Foo et al., 2017). We do not advise anybody to take this incredibly tedious approach.

Backward and forward search

Relevant published studies that are missed in main database searches can often be found with citation-based searches. After identifying pertinent reviews and landmark papers during the scoping search, you can use database records (again, Scopus and Web of Science are useful, but not the only ones; see also citationchaser (Haddaway et al., 2021) for an R package to conduct backward/forward searches) to download the reference lists of these papers (a backwards search) and their citing papers (a forward search). Forward searches are especially useful when reviewing empirical evidence for a specific hypothesis, because nearly all relevant studies cite the seminal paper which formulated that hypothesis. To minimise duplicates from multiple backward or forward searches, you can add multiple reference and citing article lists into a custom list on search platforms such as Web of Science, which will automatically remove the duplicates before you export the finalised list.

Grey literature

Grey literature refers to unpublished research (e.g., dissertations, conference abstracts, preprints, or unpublished datasets), or those published outside of traditional academic publishing (e.g., governmental reports) (see Supplementary material for more information). The inclusion of grey literature is necessary if publication bias exists in the published literature (Auger, 1998; McAuley et al., 2000). Indeed, recent studies show that the inclusion of grey literature can significantly change the conclusions of a meta-analysis (e.g., Sánchez-Tójar et al., 2018). Although grey literature can be found in major platforms such as Web of Science, or subject-specific databases, in the form of meeting abstracts, conference proceedings, or theses, it is usually more fruitful to source them from dedicated resources. At present, the major sources of quality grey literature available for ecology and evolution are PhD and Masters theses. PhD and Masters theses are located on platforms such as ProQuest Dissertation & Theses Global (covering over 3,100 institutions in over 100 countries), EBSCOhost Open Dissertations (dominated by English-speaking countries), and OpenGrey (European), which have similar Boolean search functionality as Web of Science and Scopus. Google Scholar indexes grey literature too, but they are mixed together with published articles. Google Scholar also has limited Boolean search functionality. Given these shortcomings, researchers should use Google Scholar as a supplementary rather than as the main search platform (Gusenbauer & Haddaway, 2020). With the increasing popularity of preprints and online data repositories (e.g., bioRxiv, EcoEvoRxiv, Figshare), valuable grey literature might also be found in repositories soon. In cases where the review is focusing on a relatively narrow area, it is also possible to email all/majority of researchers working in this area (it helps if they already assembled into mailing list or some specialized society) to request for unpublished data.

Removing duplicates

There are two general ‘duplicate scenarios’. First, different search methods can return the same papers. These duplicates can be removed with reference managers, literature-screening tools such as Rayyan (Ouzzani et al., 2016), or *R* packages that use simple matching or fuzzy logic (Westgate, 2019). None of these methods are fail-safe and all require user checks. Second, sister publications can duplicate a dataset (e.g., the same work presented in a thesis and a published article). In such cases, you need to manually deduplicate these studies by carefully examining full-texts of similar papers from the same group of authors and decide on rules for preferencing sources of information (e.g., choose the one with larger sample size or more transparent reporting).

The terminal investment hypothesis: searching

Identify sources and generate search terms

We chose Scopus and Web of Science as our main sources for published literature. Fig. 6 presents the process for generating our initial search string.

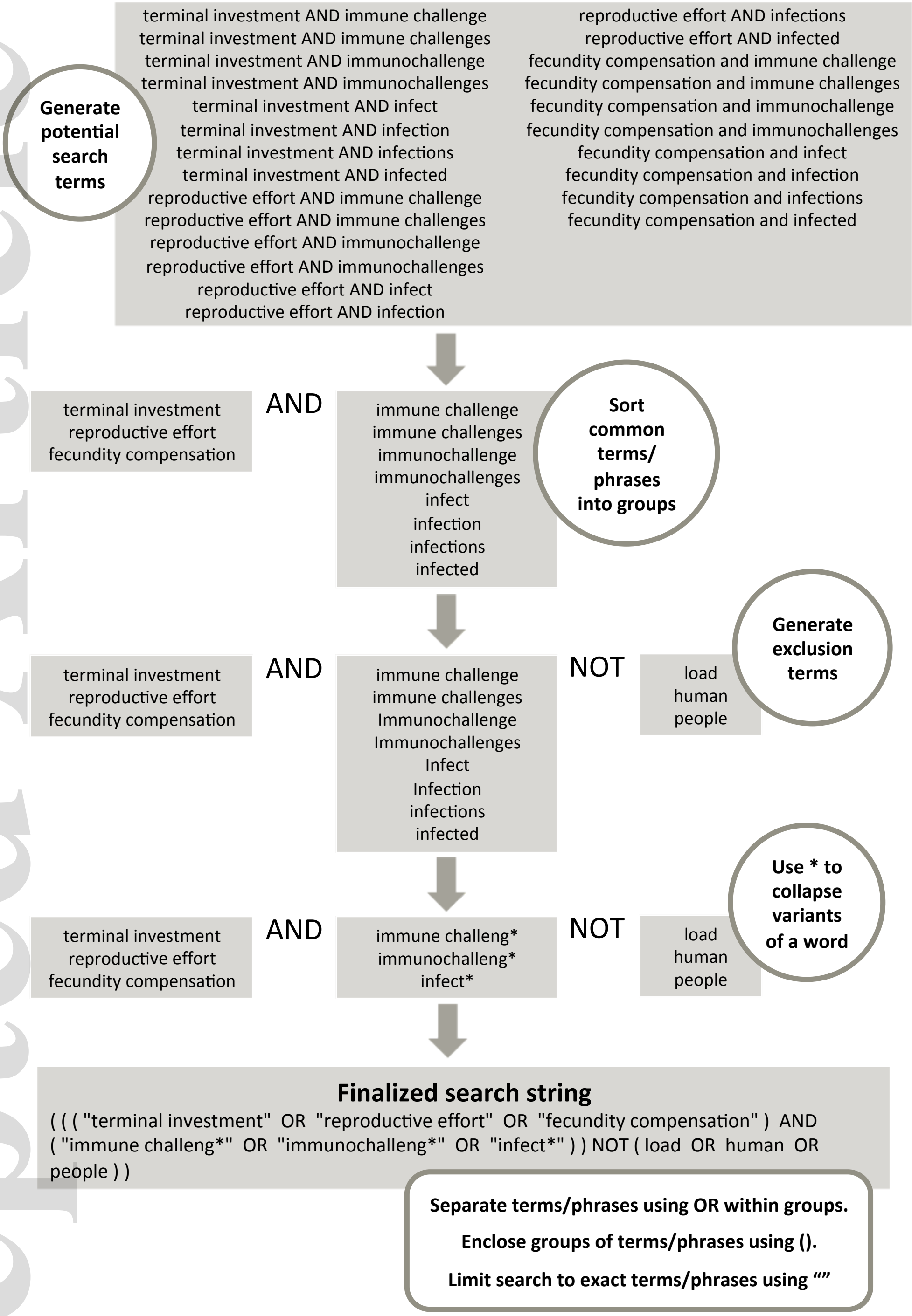

Figure 6. Generating the initial search string for the terminal investment hypothesis example. In circles / rounded rectangle we present descriptions of our main actions; grey rectangles contain actual keywords and search strings. We expected relevant research to mention words or phrases linked to reproductive investment and immune function. Thus, we generated a list of individual search strings containing both categories (e.g., reproductive effort and immune challenge). We also generated a third list of potential exclusion terms. For

instance, we limited our meta-analysis to non-human species; therefore, we aimed to exclude records that mention words such as "human" and "people". Within each group of terms, we collapsed variants of the same word using the wildcard *. Lastly, we converted the set of terms into a complete Boolean string using OR, AND, NOT, (), and "".

Refine search terms

We tested our initial search string on Web of Science (details in Supplementary Material). As shown in Fig. 7, the search string underwent multiple rounds of editing and pilot screening until we arrived at our final search string.

Stage	Search string
Initial search string (1) 159 results	***TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation") AND ("immune challeng*" OR "immunochalleng*" OR "infect*")) NOT (load OR human OR people))***
Add inclusion terms (2) 4360 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" ***OR "reproductive compensation" OR "fitness"***) AND ("immune challeng*" OR "immunochalleng*" OR "infect*" ***OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin****)) NOT (load OR human OR people))
Edit inclusion term (3) 493 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" ***OR "reproductive fitness"***) AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin*)) NOT (load OR human OR people))
Add inclusion terms (4) 2489 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" OR "reproductive fitness" ***OR "reproductive investment" OR "Life History Trade-Off*" OR "life history"***) AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin*)) NOT (load OR human OR people))
Change inclusion term (5) 1819 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" OR "reproductive fitness" OR "reproductive investment" OR "Life History Trade-Off*" ~~***OR "life history"***~~ ***OR "trade off"***) AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin*)) NOT (load OR human OR people))
Delete inclusion term (6) 1155 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" OR "reproductive fitness" OR "reproductive investment" OR "reproductive success" OR "Life History Trade-Off*" ~~***OR "trade off"***~~) AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin*)) NOT (load OR human OR people))
Add inclusion term (7) 1429 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" OR "reproductive fitness" OR "reproductive investment" OR "reproductive success" OR "Life History Trade-Off*" ***OR "Phenotypic Plasticity"***) AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin*)) NOT (load OR human OR people))
Add exclusion terms (8) 1141 results	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" OR "reproductive fitness" OR "reproductive investment" OR "reproductive success" OR "Life History Trade-Off*" OR "Phenotypic Plasticity") AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant OR vaccin*)) NOT (load OR human OR people ***OR men OR women OR infant* OR rat OR rats OR mouse OR mice OR pig* OR pork OR beef OR cattle OR sheep OR lamb* OR chicken* OR calf* OR horse****))
	Pilot 100 papers to check hit rate. 6% hit rate. Continue refining.
Final search string 1567 results (~ 10% hit rate)	TS = ((("terminal investment" OR "reproductive effort" OR "fecundity compensation" OR "reproductive compensation" OR "reproductive fitness" OR "reproductive investment" OR "reproductive success" OR "Life History Trade-Off*" OR "Phenotypic* Plastic*" OR "pre-copulatory NEAR/5 trait*" OR "sexual NEAR/5 weapon*" OR "sexual NEAR/5 ornament*" OR "post-copulatory NEAR/5 trait*" OR "ejaculate quality" OR "sperm quality" OR "mating effort" OR "parental care") AND ("immune challeng*" OR "immunochalleng*" OR "infect*" OR lipopolysaccharide OR lps OR phytohemagglutinin OR pha OR "sheep red blood cells" OR srbc OR implant* OR vaccin* OR nylon OR sephadex)) NOT (load OR human OR people OR men OR women OR infant* OR rat OR rats OR mouse OR mice OR pig* OR pork OR beef OR cattle OR sheep OR lamb* OR chicken* OR calf* OR horse* OR infective))

Figure 7. Process of refining the search string in the terminal investment hypothesis example using Web of Science search platform. The changes made at each stage are presented in taupe, bold, italic font. The search string underwent multiple rounds of editing until the total number of found records stabilised around 1,000. We then ran a pilot screening with 100

randomly selected records. The hit rate (number of relevant records) was low (6%). Therefore, we continued refining the search string until we arrived at our final search string with 1,567 records and ~10% hit rate, while retrieving 10 out of our 16 scoping search papers (Table S1). When an actual search was executed several months later, it produced 1,605 records (see Fig. 8 for the PRISMA diagram illustrating our search process and results). The same search string produced 1,478 records on Scopus (Table S2).

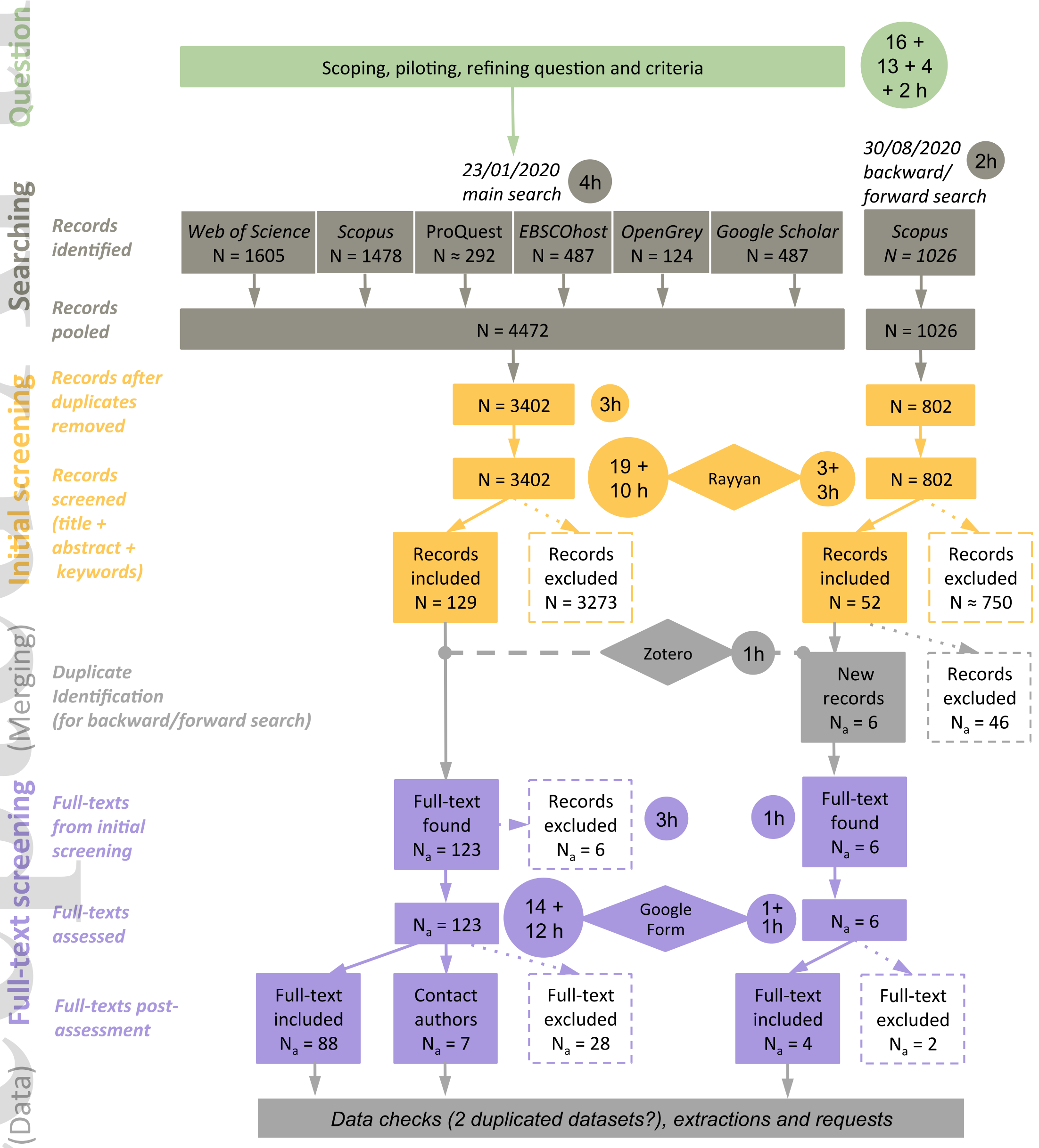

Figure 8. Modified PRISMA diagram featuring the results and time taken by each step of the terminal investment example's search process. For each step, filled rectangles represent the search records included. The unfilled rectangles with dotted outlines represent records that

were excluded. The circles represent the amount of time taken. Note that we only added the time during which we were actively executing the steps and not time in between, such as breaks. Therefore, the time between start and end of each step is likely to be longer. Also, the time reported here covers only up to the full-text screening stage. It does not include time taken for data extraction (e.g. in meta-analyses), which can take a substantial amount of time. The diamonds are the applications that we used to execute the steps.

Backward and forward search

Through our scoping search, we found the following references for our backward and forward searches: four qualitative reviews (Clutton-Brock, 1984; Duffield et al., 2017; Fischer et al., 2009; Javoiš, 2013) and the landmark paper that first used an immune challenge to test the effect of a survival challenge on reproductive investment (Bonneaud et al., 2004).

Grey literature

We searched for grey literature using ProQuest Dissertation & Theses Global, EBSCOhost Open Dissertations, Google Scholar, and OpenGrey (search details in Table S2). We found 292 records in ProQuest and 487 in EBSCOhost. For Google Scholar, we only screened the top 50 relevant results from each year (487 records in total). OpenGrey search produced 124 records for screening.

Removing duplicates

We deduplicated the records from the database searches using a free reference manager Zotero (note that some records contained errors, so the numbers do not exactly match the sum of the found references). We manually removed empty or nonsense records and used automatic deduplication in Zotero, followed by a visual check of records sorted by title. Next, we exported records into the online screening software Rayyan (described below), whose deduplication algorithm found another 13 duplicates, resulting in 3,402 unique records for screening.

We deduplicated the backward and forward search records in Scopus itself, by consolidating the records into a single automatically deduplicated list before exporting. We collected a total of 1,026 records, which resulted in 802 unique records for screening.

Try it yourself!

Conduct your own literature search with exercise 4 'Performing searches for relevant literature' in the Supplementary Material.

Initial and full-text screening

Screening typically contains two stages. First, initial screening excludes obviously irrelevant studies based on their titles, abstracts, and key words. Second, articles that are deemed potentially relevant or unclear during initial screening undergo full-text screening to confirm the list of studies that meet all inclusion criteria. Both screening stages involve: 1) generating decision trees, 2) piloting the decision trees, 3) refining the decision trees, and 4) conducting the screening.

Generating decision trees

Decision trees help researchers screen records consistently and efficiently. For example, the third question in Fig. 9a allows one to exclude records on irrelevant populations without the need to go through the rest of the questions. Full-text screening trees tend to contain more questions (Fig. 9b). Not only do researchers have to double-check the inclusion criteria from the initial screening, but also assess additional criteria that can only be judged from the full-text. For example, whether or not the research contains data required for a meta-analysis.

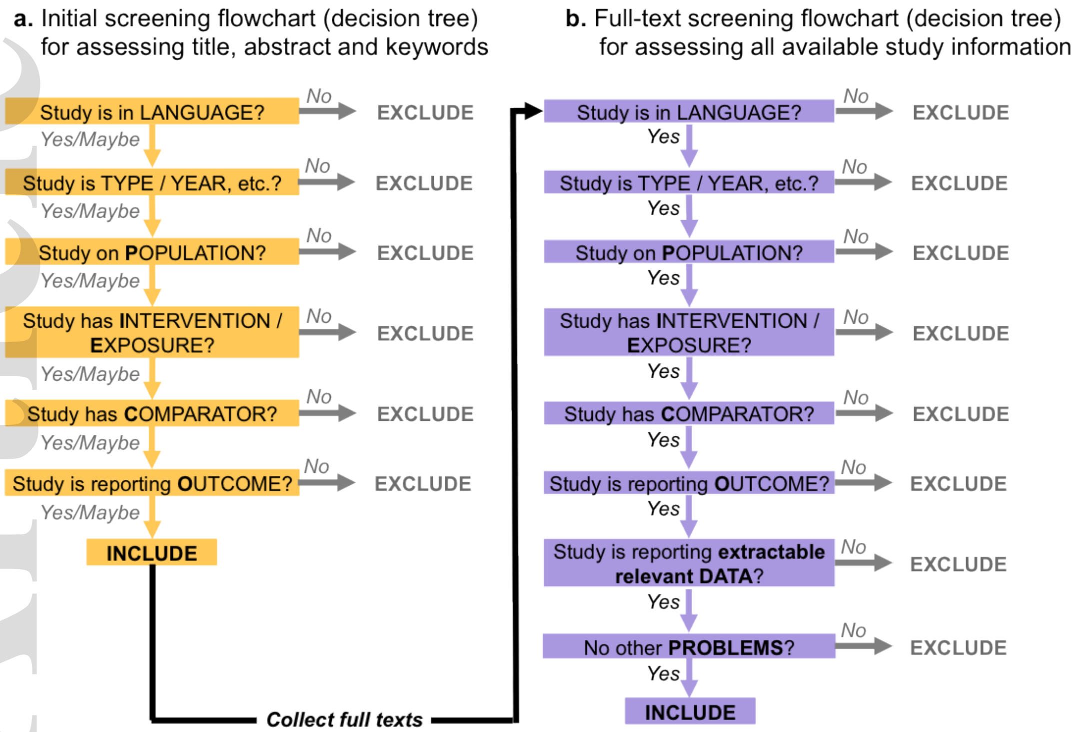

Figure 9. Abstract examples of initial (a) and full-text screening (b) decision trees. The two types differ in two ways. First, full-text screening typically contains more questions than initial screening. Second, in initial screening, records that are unclear (the 'Maybes') are retained for full-text screening, whereas in full-text screening unclear records are mostly discarded, except for those published in recent years, which are kept until authors are contacted to verify the details of the study.

Pilot and refine decision trees

Pilot screening, training, and discussions increase screening accuracy. We strongly recommend two researchers conduct pilot screening on random selections of records (at least 100 for initial and 10 for full-text screening) to ensure screening methods have at least an 80% agreement rate. Taking notes and discussing ambiguous or conflicting decisions can help refine decision trees to increase agreement rates. We recognise that having multiple people screen all records is not feasible for some projects (e.g., PhD students with limited support), but asking for help with the pilot screening is still important for increasing the objectivity of single-screened projects.

Executing screening

Reference managers are not designed for managing and tracking screening decisions. For initial screening, there are a number of specialized screening tools (see Kohl et al., 2018 for an overview). One of them is Rayyan (Ouzzani et al., 2016), which has multiple advantages, including being free and online (facilitates collaborations), allowing multiple blinded (parallel) screeners, record filtering, and allowing highlighting of screener-chosen keywords using different colours (green for inclusion and red for exclusion), for easier visual scanning of the records.

Although programs such as Rayyan can be also used for full-text screening, questionnaire programs provide good and more flexible options. Questionnaire forms can provide summaries of exclusion reasons for reporting purposes. They also allow us to begin collecting preliminary information about the papers. There is a practical limit to the number of full-texts that a small group of researchers can screen in a reasonable time. In our experience, around 3000 papers are manageable for a single screener. Downloading full-texts for screening can be time-consuming, especially if done one paper at a time. Some reference managers can facilitate this process by auto-downloading full-text files from a string of DOIs (e.g., Zotero with a plug-in).

After the full-text screening, researchers might need to contact authors for additional information that was missing or unclear from the authors' report. Typically, you are more likely to get a response for research that was made available within the last 5 years. Even then, you might not obtain a response for various reasons, including authors no longer having access to old datasets, out-dated contact information, or intention to retain priority over the requested data for future publications. We suggest sending an initial email (ccing the senior author of your team) that clearly explains the context of the planned review, identify members of the review team, the data required, and the deadline for data provision. A reminder email can be sent two weeks later, after which the study can be excluded if no response is received by a specified deadline. We provide example templates in the supplementary material for the request and reminder emails.

The terminal investment hypothesis: initial and full-text screening

Generating and piloting decision tree

We created the initial and full-text screening decision trees based on our inclusion/exclusion criteria. Two authors pilot-tested the initial screening decision tree with

100 randomly selected records from our search, and the agreement rate was high (92%). After resolving conflicting decisions, the hit-rate was 10%. The same two authors also piloted the full-text screening decision tree using 10 randomly selected records from the list that was included after the initial pilot screening, with an agreement rate of 95%. Fig. 10 presents the finalized decision trees.

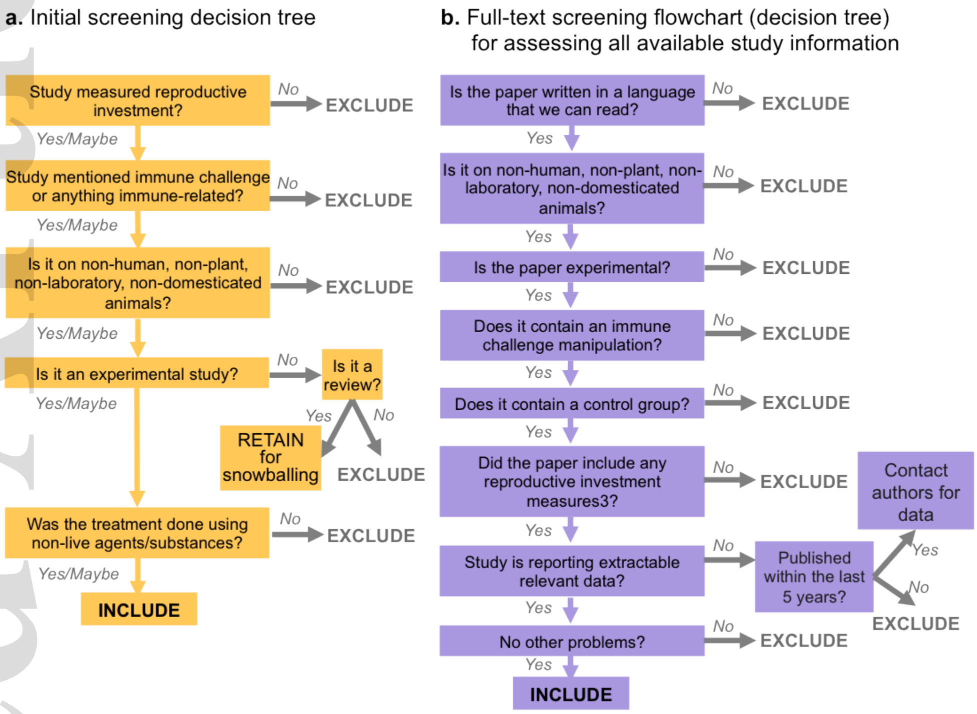

Figure 10. Refined initial (a) and full-text screening (b) decision trees for the terminal investment hypothesis meta-analysis example.

Executing screening

Two authors independently conducted the initial screening within Rayyan (Ouzzani et al., 2016). 35 conflicting decision (agreement rate of 98.9%) were discussed and resolved. We performed initial screening for the database search entries and the backward/forward search entries separately, and removed duplicated records before full-text screening. Then, the same screeners assessed 129 full texts independently using a Google form (see Supplementary Material), with an average agreement rate of 85.4%. After discussion, 99 full-texts were included (of these, 7 were flagged as requiring additional information from original authors). Finally, a manual check of the list revealed that two of the included theses were sister publications of published papers and were therefore excluded, resulting in 97 unique studies tentatively included for data extraction for meta-analysis.

Try it yourself!

Create decision trees and conduct some initial screening with exercises 3 'Making decision trees for the literature screening' and 5 'Screening the literature' in the Supplementary Material.

Discussion

We presented a practical guide on conducting systematic searches for ecologists and evolutionary biologists, focusing on practical recommendations for beginners. In Fig. 11, we have compiled a list of the tips and tricks covered in this paper. Although we provided concrete numbers for some sub-steps, these are just "rules of thumb", not prescriptive benchmarks. For those seeking more detailed guidance on the methods of systematic searches, there is a large collection of resources available, including free-access online library guides, published articles (e.g., Bartels, 2013; Gusenbauer & Haddaway, 2020; see also https://vortal.htai.org/?q=sure-info for a set of resources), handbooks on systematic reviews (e.g., Koricheva et al., 2013), resources from prominent systematic review networks and consortiums (Campbell Collaboration, Cochrane, and Collaboration for Environmental Evidence), and personal assistance from information specialists/librarians. Authors can also check the quality of their searches using appraisal tools such as PRISMA Eco-Evo (O'Dea et al., accepted), AMSTAR (Shea et al., 2017), and CEESAT (Woodcock et al., 2014).

Figure 11. Summary of tips and tricks to conducting a systematic search.

Registration

Recent papers have highlighted a lack of reproducibility in the reporting and results of meta-analyses across fields (Davies et al., 2020; Jones & DuVal, 2019; Koricheva & Gurevitch, 2014; Lakens et al., 2017). The reason could partially be attributable to the lack of adherence to reporting standards (e.g., PRISMA; Moher, Shamseer, et al., 2015). As part of the credibility revolution, there is a growing emphasis on registering one's study (Nosek et al., 2018; Parker et al., 2019). There are three main formats currently available: pre-registration (no requirements on the level of detail), registered reports (with peer-review of introduction and methods for publication in journals before the research is conducted), and published protocols (similar to registered reports, except that the protocol is published separately from the final review, e.g., see Cochrane systematic reviews). A number of journals many ecologists and evolutionary biologists will be familiar with publish registered reports, including *eLife*, *BMC Biology*, *PLoS Biology*, *Ecology and Evolution, Conservation Biology,* and *Ecological Solutions and Evidence*, and the journals *Environmental Evidence* and *Systematic Reviews* publish protocols of systematic reviews and maps.

The terminal investment meta-analysis example reported in this paper is a Stage-1-accepted registered report at *BMC Biology*. It is one of the first meta-analytic registered reports in the field of ecology and evolution. Having produced this registered report, we have four recommendations on the preparation of pre-registrations or registered reports. First, be as explicit as possible about all the steps (listed in Fig. 1). Second, pilot test wherever possible to verify proposed steps and report them accordingly (e.g., question scope, search strings, decision trees). Third, some decisions have to be made after collecting the papers. Therefore, there may be more than one possible path to take. In such cases, note down all the possible forking paths of decisions in the methods. Finally, document the review steps obsessively, especially information required by reporting standards (e.g., PRISMA).

Concluding remarks

A systematic search is typically associated more with systematic reviews and meta-analyses than with other forms of reviews, such as narrative reviews or opinion articles. However, reviews, regardless of type, can accurately reflect the state of understanding of a field only when they are based on a representative set of papers. Therefore, we advocate that all reviews should be conducted systematically. Furthermore, we believe that a systematic search approach can be extremely useful to primary researchers too. Primary researchers can use this approach, or at least some of the principles we described, to accurately inform their introduction and discussion sections. Overall, we hope this paper will encourage more researchers, both synthesists and primary researchers, to adopt a systematic and reproducible approach when searching the literature.

Acknowledgement

We thank Miss Allie Ofisher for creating the colour-blind friendly colour palette (https://medium.com/cafe-pixo/inclusive-color-palettes-for-the-web-bbfe8cf2410e) that we used for the figures in this paper.

Funding

YZF is supported by an Australian Government Endeavour Fellowship (ERF_Pdr_6537_2018), ML and SN are supported by a Discovery grant (DP180100818). The funders had no role in developing this protocol.

Competing interests

None of the authors have any competing interests.

Author contribution

YZF, ML, SN, and REO conceptualised the paper. Y.Z.F drafted the manuscript. ML drafted the supplementary material. All authors provided critical revisions and approved the manuscript for submission.

Data Availability

No datasets were generated for this paper.

References

Auger, C. P. (1998). *Information Sources in Grey Literature (Guides to Information Sources).* (4th Editio). Bowker Saur.

Bar-Ilan, J. (2018). Tale of Three Databases: The Implication of Coverage Demonstrated for a Sample Query. *Frontiers in Research Metrics and Analytics*, *3*(February). https://doi.org/10.3389/frma.2018.00006

Bartels, E. M. (2013). How to perform a systematic search. *Best Practice and Research Clinical Rheumatology*, *27*(2), 295–306. https://doi.org/10.1016/j.berh.2013.02.001

Bonneaud, C., Mazuc, J., Chastel, O., & Westerdahl, H. (2004). Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. *Evolution*, *58*(12), 2823–2830. papers3://publication/uuid/36ED4FA1-B917-4533-9633-F463E2CFF584

Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Systematic Reviews*, *1*(1), 2. https://doi.org/10.1186/2046-4053-1-2

Clutton-Brock, T. H. (1984). Reproductive effort and terminal investment in iteroparous animals. *The American Naturalist*, *123*(2), 212–229.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, *62*(7), 1382–1402. https://doi.org/10.1002/asi.21525

Davies, A. D., Lewis, Z., & Dougherty, L. R. (2020). A meta-analysis of factors influencing the strength of mate-choice copying in animals. *Behavioral Ecology*, *31*(6), 1279–1290. https://doi.org/10.1093/beheco/araa064

Duffield, K. R., Bowers, E. K., Sakaluk, S. K., & Sadd, B. M. (2017). A dynamic threshold model for terminal investment. *Behavioral Ecology and Sociobiology*, *71*, 185. https://doi.org/10.1007/s00265-017-2416-z

Fellows, I. (2018). *wordcloud: Word Clouds* (2.6). R package.

Fischer, B., Taborsky, B., & Dieckmann, U. (2009). Unexpected patterns of plastic energy allocation in stochastic environments. *American Naturalist*, *173*(3). https://doi.org/10.1086/596536

Fisher, R. (1930). *The genetical theory of natural selection*. Clarendon Press. https://doi.org/10.1111/jeb.12566

Foo, Y. Z., Nakagawa, S., Rhodes, G., & Simmons, L. W. (2017). The effects of sex hormones on immune function: a meta-analysis. *BIOLOGICAL REVIEWS*, *92*(1), 551–571. https://doi.org/10.1111/brv.12243

Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, *10*(10), 1645–1654. https://doi.org/10.1111/2041-210X.13268

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

Haddaway, N. R., Grainger, M. J., & Gray, C. T. (2021). *citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching*. https://doi.org/10.5281/zenodo.4543513

Haddaway, N. R., Kohl, C., Rebelo Da Silva, N., Schiemann, J., Spök, A., Stewart, R., Sweet, J. B., & Wilhelm, R. (2017). A framework for stakeholder engagement during systematic reviews and maps in environmental management. *Environmental Evidence*, *6*(1). https://doi.org/10.1186/s13750-017-0089-8

Hjørland, B. (2014). Classical databases and knowledge organization: A case for boolean retrieval and human decision-making during searches. *Journal of the Association for Information Science and Technology*, *66*(8), 1559–1575. https://doi.org/10.1002/asi.23250

Javoiš, J. (2013). A two-resource model of terminal investment. *Theory in Biosciences*, *132*(2), 123–132. https://doi.org/10.1007/s12064-013-0176-5

Jones, B. C., & DuVal, E. H. (2019). Mechanisms of Social Influence: A Meta-Analysis of the Effects of Social Information on Female Mate Choice Decisions. *Frontiers in Ecology and Evolution*, *7*(October), 1–14. https://doi.org/10.3389/fevo.2019.00390

Khan, K., Kunz, R., Kleijnen, J., & Antes, G. (2011). *Systematic reviews to support evidence-based medicine* (2nd ed.). Crc Press.

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), 1–17. https://doi.org/10.1186/s13750-018-0115-5

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, *102*(4), 828–844. https://doi.org/10.1111/1365-2745.12224

Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). Handbook of meta-analysis in ecology and evolution. In J. Koricheva, J. Gurevitch, & K. Mengersen (Eds.), *Handbook of Meta-analysis in Ecology and Evolution*. Princeton University Press. https://doi.org/10.23943/princeton/9780691137285.001.0001

Lakens, D., Tilburg, M. van A., Anvari, F., Corker, K. S., Grange, J. A., Gerger, H., Hasselman, F., Koyama, J., Locher, C., Miller, I., Page-Gould, E., Schönbrodt, F. D., Sharples, A., Spellman, B. A., & Zhou, S. (2017). Examining the reproducibility of meta-analyses in Psychology: A preliminary report. *MetaArXiv*. https://doi.org/10.1017/CBO9781107415324.004

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. In *Practical meta-analysis.* Sage Publications, Inc.

McAuley, L., Ba'Pham, Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet*, *356*(9237), 1228–1231. https://doi.org/10.1016/S0140-6736(00)02786-0

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Estarli, M., Barrera, E. S. A., Martínez-Rodríguez, R., Baladia, E., Agüero, S. D., Camacho, S., Buhring, K., Herrero-López, A., Gil-González, D. M., Altman, D. G., Booth, A., … Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*, 1. https://doi.org/10.1186/2046-4053-4-1

Moher, D., Stewart, L., & Shekelle, P. (2015). All in the family: Systematic reviews, rapid reviews, scoping reviews, realist reviews , and more. *Systematic Reviews*, *4*, 183. https://doi.org/10.1186/s13643-015-0163-7

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., & Nakagawa, S. (n.d.). Preferred Reporting Items for Systematic reviews and Meta-Analyses in ecology and evolutionary biology: a PRISMA extension. *Biological Reviews*.

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

Parker, T., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, *33*(4), 747–750. https://doi.org/10.1111/cobi.13342

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *Acp Journal Club*, *123*(3), A12-3.

Rivest, M. (2016). *scimeetr: Scientometrics Meet R*.

Roberts, M. L., Buchanan, K. L., & Evans, M. R. (2004). Testing the immunocompetence handicap hypothesis: A review of the evidence. *Animal Behaviour*, *68*(2), 227–239. https://doi.org/10.1016/j.anbehav.2004.05.001

Sánchez-Tójar, A., Nakagawa, S., Sánchez-Fortún, M., Martin, D. A., Ramani, S., Girndt, A., Bókony, V., Kempenaers, B., Liker, A., Westneat, D. F., Burke, T., & Schroeder, J. (2018). Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. *ELife*, 1–26. https://doi.org/10.1101/283150

Santos, E. S. A., & Nakagawa, S. (2012). The costs of parental care: A meta-analysis of the trade-off between parental effort and survival in birds. *Journal of Evolutionary Biology*, *25*(9), 1911–1917. https://doi.org/10.1111/j.1420-9101.2012.02569.x

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ (Online)*, *358*, 1–9. https://doi.org/10.1136/bmj.j4008

Vico, G., Manzoni, S., Nkurunziza, L., Murphy, K., & Weih, M. (2016). Trade-offs between seed output and life span - a quantitative comparison of traits between annual and perennial congeneric species. *New Phytologist*, *209*(1), 104–114. https://doi.org/10.1111/nph.13574

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

Woodcock, P., Pullin, A. S., & Kaiser, M. J. (2014). Evaluating and improving the reliability of evidence syntheses in conservation and environmental science: A methodology. *Biological Conservation*, *176*, 54–62. https://doi.org/10.1016/j.biocon.2014.04.020

Zupic, I., & Čater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods*, *18*(3), 429–472. https://doi.org/10.1177/1094428114562629