Type	Number If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_ Supplementary Video 1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Table	1	Haddaway et al. Supplementary File.xlsx	Supplementary Table 1. Examples of literature reviews and common problems identified.

Eight problems with literature reviews and how to fix them

Neal R Haddaway[1,2,3*], Alison Bethel[4], Lynn V. Dicks[5], Julia Koricheva[6], Biljana Macura[1], Gillian Petrokofsky[7], Andrew S Pullin[8], Sini Savilaakso[9,10], Gavin B Stewart[11]

Article impact statement: Systematic reviews can easily fall foul of eight key pitfalls commonly found in poor reviews. However, these pitfalls can be readily avoided.

Running head: The road to reliable systematic reviews

Keywords: decision-support, evidence-based conservation, evidence-informed decision-making, evidence synthesis, research waste

Word count: 5,192

[1] Mercator Research Institute on Climate Change and Global Commons, Torgauer Str. 19, 10829 Berlin, Germany
[2] Stockholm Environment Institute, Linnégatan 87D, 10451 Stockholm, Sweden
[3] Africa Centre for Evidence, University of Johannesburg, 1 Ibis St, Cottesloe, Johannesburg, 2092, South Africa
[4] College of Medicine and Health, Exeter University, Exeter, EX1 2LU, UK
[5] Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK; School of Biological Sciences University of East Anglia, Norwich NR4 7TJ, UK
[6] Department of Biological Sciences, Royal Holloway University of London, Egham, Surrey, TW20 0EX, UK
[7] Department of Zoology, University of Oxford, Oxford, OX1 3SZ, UK
[8] Collaboration for Environmental Evidence, UK Centre, School of Natural Sciences, Bangor University, Bangor, LL57 2UW, UK
[9] Liljus ltd, 3rd Floor, 7 Moorgate, London, United Kingdom, EC2R 6AF, UK
[10] University of Helsinki, Department of Forest Sciences, P.O. Box 4, (Yliopistonkatu 3), 00014 University of Helsinki, Finland
[11] Evidence Synthesis Lab, School of Natural and Environmental Sciences, University of Newcastle, Newcastle-upon-Tyne, NE1 7RU, UK

Corresponding author: neal_haddaway@hotmail.com; neal.haddaway@sei.org

Eight problems with literature reviews and how to fix them

Abstract

Traditional approaches to reviewing literature may be susceptible to bias and result in incorrect decisions. This is of particular concern when reviews address policy- and practice- relevant questions. Systematic reviews have been introduced as a more rigorous approach to synthesising evidence across studies; they rely on a suite of evidence-based methods aimed at maximising rigour and minimising susceptibility to bias. Despite the increasing popularity of systematic reviews in the environmental field, evidence synthesis methods continue to be poorly applied in practice, resulting in the publication of syntheses that are highly susceptible to bias. Recognising the constraints that researchers can sometimes feel when attempting to plan, conduct and publish rigorous and comprehensive evidence syntheses, we aim here to identify major pitfalls in the conduct and reporting of systematic reviews, making use of recent examples from across the field. Adopting a 'critical friend' role in supporting would-be systematic reviews and avoiding individual responses to police use of the 'systematic review' label, we go on to identify methodological solutions to mitigate these pitfalls. We then highlight existing support available to avoid these issues and call on the entire community, including systematic review specialists, to work towards better evidence syntheses for better evidence and better decisions.

Background

The aims of literature reviews range from providing a primer for the uninitiated to summarising the evidence for decision making [1]. Traditional approaches to literature reviews are susceptible to bias and may result in incorrect decisions [2, 3]. This can be particularly problematic when reviews address applied, policy-relevant questions, such as human impact on the environment or effectiveness of interventions where there is a need for review results to provide a high level of credibility, accountability, transparency, objectivity, or where there is a large or disparate evidence base or controversy and disagreement amongst existing studies. Instead, rigorous approaches to synthesising evidence across studies may be needed, i.e. systematic reviews.

Systematic review is a type of research synthesis that relies on a suite of evidence-based methods aimed at maximising rigour and minimising susceptibility to bias. This is achieved by attempting to increase comprehensiveness, transparency, and procedural objectivity of the review process [4]. The methods involved are outlined in Figure 1 [see also 2, 5].

Systematic reviews were originally developed in the fields of social science and healthcare and have had a transformative effect, particularly in health, where they underpin evidence-based medicine [6]. Introduction of systematic reviews into medicine was facilitated by Cochrane, the review coordinating body that sets standards and guidance for systematic reviews of healthcare interventions (https://www.cochrane.org/). Systematic reviews are now increasingly published in other fields, with the Collaboration for Environmental Evidence (CEE) established in 2008 to act as the coordinating body supporting efforts in the field of conservation and environmental management (see http://www.environmentalevidence.org).

Towards a better understanding of rigour in evidence synthesis

Despite the increasing popularity of systematic reviews in the environmental field, evidence synthesis methods continue to be poorly applied in practice, resulting in the publication of syntheses that are highly susceptible to bias. In one assessment by O'Leary et al. [7], a set of 92 environmental reviews published in 2015 was judged to be poorly conducted and reported (a median score of 2.5 out of a possible 39 using the synthesis appraisal tool CEESAT, Woodcock et al. [8]). Substandard reviews could provide misleading findings, potentially causing harm and wasting valuable resources in research, policy and practice. Furthermore, these reviews could erode trust in evidence synthesis as an academic endeavour.

Substantial support exists to help raise the rigour of evidence synthesis toward the recognised standards of systematic reviews: a range of Open Access methodological guidance and standards exists both across subjects [9, 10] and in the field of conservation and environment [5]. Methods for peer-reviewing and critically appraising the rigour of systematic reviews are also freely available [8, 11]. Open Educational resources in evidence synthesis methodology exist online (e.g. [1] and https://synthesistraining.github.io/). There are free-to-use, online platforms designed to support the methodology, such as SysRev (https://sysrev.com). Finally, an open and willing community of practice consisting of hundreds of methodologists exists in the field of conservation and environment (CEE, www.environmentalevidence.org), as it does in social policy (the Campbell Collaboration, www.campbellcollaboration.org) and healthcare (Cochrane, www.cochrane.org). That said, the lack of awareness and adherence to internationally accepted minimum standards and best practices in evidence synthesis in the field of conservation and environment demonstrates that more must be done to support the publication of reliable syntheses. Despite all these clear international standards and freely accessible and abundant guidance for systematic reviews, review articles are frequently published that claim to be 'systematic reviews', because they have employed some elements of the method, but fall substantially short of the standard [12]. In sum, we see two related issues when considering rigour of evidence syntheses. Firstly, most published evidence reviews are poorly conducted. Secondly, those that describe themselves as 'systematic reviews' imply an increased level of rigour, and where this is not true (i.e. the authors have failed to adequately follow accepted standards), confusion occurs over what the term 'systematic review' really means.

Here, we describe issues affecting all evidence reviews and encourage review authors to transparently report their methods so that the reader can judge how systematic they have been. We do not believe that all reviews should be 'systematic reviews'; for example, 'primers' or overviews to a novel topic or reviews that combine concepts do not seek to be comprehensive, rigorous or definitive in influencing policy. However, we do believe that all reviews can benefit from applying some of these best practices in systematic approaches, with transparency perhaps being the least costly to operationalise.

We understand the resource and time constraints faced by review authors, and we appreciate the costs involved in attempting to produce and publish rigorous evidence syntheses. However, we do believe that the reliability of reviews intended to inform policy is a serious scientific and social issue and could be substantially improved if the research community were to fully embrace rigorous evidence synthesis methods, committing to raise awareness across the board. We also know that this can be achieved incrementally, progressively increasing the standard of reviews produced over time, and without necessarily breaking the bank when it comes to resources and funding.

Objectives

Recognising the constraints that researchers can sometimes face when attempting to plan, conduct and publish rigorous and comprehensive evidence syntheses, we aim here to identify major pitfalls in the conduct and reporting of systematic reviews, making use of recent examples from across the field. Adopting a 'critical friend' role of supporting potential systematic reviewers, we go on to identify methodological solutions to mitigate these pitfalls. We then highlight existing support available to avoid these issues. Finally, we describe key intervention points where the conservation and environmental management communities, including funders, review authors, editors, peer-reviewers, educators, and us as methodologists, can act to avoid problems associated with unreliable and substandard reviews.

8 problems, 8 solutions

In the following section, we use recent examples of literature reviews published in the field of conservation and environmental science to highlight 8 major limitations and sources of bias in evidence synthesis that undermine reliability. We describe each problem and provide potential mitigation solutions in turn. The problems, examples and solutions for different actors are outlined in Supplementary Information.

1. *Lack of relevance (limited stakeholder engagement)*

Description: Taking a broad definition of stakeholders (including any individual or group who is affected by or may affect the review and its findings [13]), all reviews whose results will be used either to shape an academic field or to inform policy or practice decision making should involve some degree of stakeholder engagement. Doing so can improve review effectiveness, efficiency and impact [14, 15]. In some 'public goods' reviews (i.e. those published and intended for a wide audience [16]), however, authors do not adequately engage with relevant stakeholders. This may result in the scope of the review being of limited practical relevance to researchers and decision-makers. It may also result in the review using definitions of key concepts and search terms that are not broadly accepted or appropriate, limiting acceptance and uptake of the review's findings, or producing an inaccurate or biased selection of literature. This may result from a lack of coherence within the stakeholder communities themselves. Stakeholder engagement in evidence synthesis is an opportunity for attempting to resolve these issues, however; providing broad benefits to the wider science-policy and -practice community.

Example: In conducting the systematic review on the impacts of palm oil production on biodiversity, Savilaakso et al. [17] contacted recognised experts and key stakeholders as outlined in the protocol [18]. Although the authors contacted company representatives, in retrospect the stakeholder engagement was not broad enough. After publication of the review, the Malaysian palm oil industry criticised the review for its narrow focus on biodiversity and not including poverty impacts. A broader stakeholder engagement could have alleviated the problem by explaining the purpose of the review (i.e. review of existing knowledge as a starting point for research proposals related to land-use) and/or it could have led to a broader review inclusive of social impacts.

Mitigation strategies: Stakeholder engagement can require substantial resources if reviewers aim for it to be comprehensive and include physical meetings, particularly on contentious topics. However, stakeholders can readily be identified, mapped and contacted for feedback and inclusion without the need for extensive budgets. Reviewers could, as a minimum, attempt to identify important minorities or marginalised groups and then engage with key groups remotely, asking for feedback on a brief summary of the planned review by email [14, 19]. This should be described in the review report.

2. *Mission creep and lack of a protocol*

Description: Mission creep occurs when the review deviates from the initial objectives. Key definitions, search strategies and inclusion or appraisal criteria may alter over time or differ between reviewers. The resultant set of articles will then not be representative of the relevant evidence base and important studies may have been omitted. As a result, the review may be highly inaccurate and misleading, and will be unrepeatable. *A priori* protocols minimise bias, allow constructive feedback before mistakes in review methodology are made, allow readers to verify methods and reporting, and act as a within-group roadmap in methods during conduct of the review. Reviews that lack protocols preclude this clarity and verifiability. This is similar to 'pre-registering' of primary research in some fields, where methodological plans are published, date-stamped, versioned and are unalterable).

Example: In their review of insect declines, Sánchez-Bayo and Wyckhuys [20] failed to provide a protocol and succumbed to mission creep. They did so by initially focusing on drivers of insect decline as described in the objectives, but shifting to generalise about insect populations across all species, not just those declining. Their searches focused exclusively on studies identifying declining populations, but their conclusions purportedly relate to all insect populations. Similarly, Agarwala and Ginsberg [21] reviewed the tragedy of the commons and common-property resources but failed to provide a protocol that would justify the choice of search terms and clarify the criteria selecting studies for the review.

Mitigation strategies: Review authors should carefully design an *a priori* protocol that outlines planned methods for searching, screening, data extraction, critical appraisal and synthesis in detail. This should ideally be peer-reviewed and published (journals such as Environmental Evidence, Ecological Solutions and Evidence, and Conservation Biology now accept registered reports/protocols, and protocols can be stored publicly on preprint servers such as Open Science Framework Preprints [https://osf.io/preprints]), and may benefit substantially from stakeholder feedback (see point 1 above). Occasionally, deviations from the protocol are necessary as evidence emerges, and these must be detailed and justified in the final report.

3. *Lack of transparency/replicability (inability to repeat the study)*

Description: An ability to repeat a review's methods exactly (also referred to as 'replicability') is a central tenet of the scientific method [22], and the methods used to produce reviews should be reported transparently in sufficient detail to allow the review to be replicated or verified [23]. If the reader can understand neither how studies were identified, selected and synthesised, nor which were excluded, the risk of bias cannot be assessed, and unclear subjective decisions may affect reliability. Unreplicable reviews cannot truly be trusted, since mistakes may have been made during conduct. In addition, unreplicable reviews have limited legacy, since they cannot be upgraded or updated and differences in outcomes between several reviews on the same topic cannot be reconciled. Ultimately, unreplicable reviews erode trust in evidence synthesis as a discipline, creating a barrier to evidence-informed policy. Similarly, a lack of transparency in reporting what was found (i.e. raw study data, summary statistics, and analytical code) prevents analytical replication and verification.

Example: Lwasa et al. [24], in their review of the mediating impacts of urban agriculture and forestry on climate change, failed to describe their methods in sufficient detail; for example, which grey literature sources and which databases/indexes within Web of Science were searched. In addition, the authors reported only some of the terms that were included in the bibliographic searches. In their review of the impact of species traits on responses to climate change, Pacifici et al. [25] did not describe how their inclusion criteria were applied in practice, so it is impossible to know whether or how they dealt with subjectivity and inconsistency between reviewers. More problematic, Owen-Smith [26] and Prugh et al. [27] failed to include a methods section of any kind in their reviews. Also problematic, and perhaps more common than a failure to describe methods, is a failure to include the extracted data. For example, Li et al. [28] did not present their data, which prevents replication of their analyses or later updating of their synthesis.

Mitigation strategies: Making use of high-standard evidence syntheses and guidance (such as those published by Cochrane, the Campbell Collaboration and CEE) as examples can help improve reporting. Similarly, review authors should attempt to conform to internationally accepted review reporting standards, such as PRISMA [29] and ROSES [23], to ensure all relevant methodological information has been included in protocols and review reports. Additionally, review authors can choose to include methodology experts in their review teams or advisory groups. Finally, review authors can choose to publish their syntheses through leading organisations and journals working with systematic reviews and maps, such as CEE.

Review authors should provide meta-data (descriptive information), data (individual study findings), and analytical code (e.g. R scripts used for meta-analysis) in full alongside their review as far as is legally permitted, and summary data where not. Guidelines (https://data.research.cornell.edu/content/writing-metadata) and example systematic reviews [e.g. 30] can highlight best practices in meta-data creation. Where authors' decisions are known to be somewhat subjective, for example on issues relating to study validity, review authors should first trial assessments and then discuss among co-authors all inconsistencies in detail before continuing. In addition, reviewers should report in detail all decisions, for example: which studies are eligible, what data should be extracted, and how valid studies are viewed to be, along with justifications for these decisions. This then allows actions to be fully understood and replicated.

4. *Selection bias and a lack of comprehensiveness (inappropriate search methods and strategy)*

Description: Selection bias occurs where the articles included in a review are not representative of the evidence base as a whole [31]. Any resultant synthesis and conclusions based on this evidence are then highly likely to be biased or inaccurate. Broadly speaking, selection bias may occur in reviews as a result of failing to account for bias in what research is published (publication bias) and what data are reported in published studies (reporting bias), and by substandard review methods that affect which studies are included in the review. Specifically in relation to search strategies, however, selection bias affects syntheses through inappropriate search strategies; for example, as a result of 'cherry picking' studies for inclusion, choosing biased/unrepresentative bibliographic databases, or using inappropriate search strategies for the subject at hand.

Example: By including 'decline' as a search term, Sánchez-Bayo and Wyckhuys [20] targeted only studies showing a reduction in insect population, contradicting their goal to collate "all long-term insect surveys conducted over the past 40 years". Thus, the authors synthesised a subset of evidence based on the direction of observed results, potentially missing studies showing a neutral or positive change, and exaggerating the insect populations' declining status. Furthermore, the authors' search was not comprehensive, including no synonyms, which are vital to account for differences in how researchers describe a concept. Their string will have missed any research using other terms that may be important synonyms; for example, 'reduction' as well as 'decline'. Adding the term 'increas*' would retrieve a significant additional body of evidence. Secondly, the review authors searched only one resource, Web of Science (they probably mean Web of Science Core Collections, but the exact indexes involved would still be unclear). The authors also excluded/ignored grey literature (see point 5, below).

In a review of tropical forest management impacts [32] and in a review of forest conservation policies [33] searches for evidence were performed only within Google Scholar, relying on Google's relevance-based sorting algorithm that displays only the first 1,000 records, which likely provides a biased subset of the literature and has been widely shown to be inappropriate as a main source of studies for literature review [34-36].

Mitigation strategies: Search methods should include more than bibliographic database searching; supplementary methods should also be employed, for example forwards and backwards citation searching, web searching, and calls for submission of evidence. Search strategies should be carefully planned and should include a comprehensive set of synonyms relevant to the review scope. Specifically, the strategy should: 1) be based on thorough scoping of the literature; 2) be trialled in a sample database and tested to ensure it recovers studies of known relevance (benchmarking [37]); 3) should ideally be constructed by or with input/support from an information specialist/librarian; 4) involve searches of multiple bibliographic databases (ranging in subject/geographic/temporal scope; for example Scopus, CAB Abstracts and MEDLINE) to maximise comprehensiveness and mitigate bias; and 6) be outlined in an *a priori* protocol that is published and open for scrutiny.

5. *Publication bias (exclusion of grey literature and failure to test for evidence of publication bias)*

Description: This issue is closely related to and perhaps a subset of Problem 4 above, but nevertheless requires a separate discussion due to the nature of the mitigation strategies necessary. Positive and statistically significant research findings are more likely to be published than negative and non-significant results [38]. The findings of syntheses based only on traditional, commercially published academic research will be as biased as the underlying research. Research that is not published in traditional academic journals controlled by commercial publishers is called 'grey literature', and consists of two main groups - the 'file-drawer' research that was intended to be published in an academic outlet but for some reason was not; where this reason was a lack of statistical or perceived biological significance, publication bias has occurred. A second type of grey literature consists of organisational reports and other studies that were not intended for an academic audience. Where relevant studies of this type are omitted from a review, the evidence base will lack comprehensiveness (see point 4 above). Tests that lead one to strongly suspect the presence of publication bias and/or quantify its potential impact are an important element of a high-quality quantitative synthesis (Egger Test, Vivea and Hedges tests [39]).

Example: In their recent review, Agarwala and Ginsberg [21] ignored grey (i.e. not commercially published) literature, excluding organisational reports and theses shown to be valuable sources of evidence [30]. When the authors then critically appraised studies, there was no justification for avoiding grey literature on the grounds of validity, and including it could have reduced the probability of publication bias. Pacifici et al. [25] also failed to include grey literature. As a result, the included evidence is likely to be unreliable (although their summaries are arguably more dangerous because of vote-counting (see point 7, below).

Mitigation strategies: Review authors should attempt to identify and include relevant grey literature in their syntheses [37, 40]. This can be attempted by searching specifically for file-drawer research in thesis repositories and catalogues, preprint servers, and funders' registries. Calls can also be made for researchers to submit unpublished studies. Organisational reports should be searched for by screening websites and physical repositories of relevant organisations, and by searching on specific bibliographic databases or web-based academic search engines, such as Google Scholar. Review authors should attempt to identify publication bias in their syntheses by conducting appropriate tests (e.g. Egger test) and visualisations (e.g. funnel plots) that may suggest publication bias as a feasible reason for heterogeneity between large and small studies [41].

6. *Lack of appropriate critical appraisal (treating all evidence as equally valid)*

Description: Some primary research is less reliable than others because of problems with the methods used, potentially resulting in an inaccurate or biased finding [42]. Reviews that fail to appropriately assess and account for the reliability of included studies are susceptible to perpetuating these problems through the synthesis, resulting in inaccurate and biased findings. Primary research may have issues relating to 'internal validity' (i.e. the accuracy of methods) that are caused, for example, by confounding variables, a lack of blinding, failure to account for the presence of confounding variables, and a lack of randomisation. Reviews may also suffer from problems with external validity, whereby primary studies vary in their relevance to the review question (for example being conducted across different spatial scales) but this is not accounted for in the synthesis. Finally, review conclusions may be misleading if studies are selected for meta-analysis based on criteria that do not properly relate to the study question.

Englund et al. [43] provide an illustrative example of how criteria influence study selection and subsequent meta-analysis results. Their datasets on stream predation experiments vary from all-inclusive criteria to minimal subset of studies. The study shows how meta-analytic patterns can appear and disappear based on the selection criteria applied.

Example: Burivalova et al. [32] included in their review a variety of studies from meta-analysis to case studies. Their stated goal was "to compare forest variables under two different management regimes, or before and after management implementation" in tropical forests. They did not conduct critical appraisal of the studies and ended up including studies that lacked either internal or external validity. For example, they included an earlier study by Burivalova et al. [44] that looked at the importance of logging intensity as a driver of biodiversity decline in timber estates. However, conclusions about logging intensity were hampered by a failure to consider log extraction techniques, and this failure had already been noted by Bicknell et al. [45] who sought to account for the influence of extraction techniques with meta-analysis. Burivalova et al. [32] also included a study by Damette and Delacote [46] that used global country-level data to study deforestation and assess sustainability of forest harvesting. Although some of the results were given separately for developing countries, the dataset used to assess certification impacts included countries globally and thus lacked external validity in a review focused on tropical forests only. Similarly, they included a study by Blomley et al. [47] that compared participatory forest management to government managed forests in Tanzania without reporting any baseline differences or matching criteria for the different forest areas.

Mitigation strategies: Systematic reviews should include a critical appraisal of every included study's internal and external validity [5]. This assessment should be carefully planned *a priori* and trialled to ensure that it is fit-for-purpose and that review authors can conduct the appraisal consistently [10]. Existing critical appraisal tools used in other reviews may prove a useful starter from which to develop a suitable tool [42]. Critical appraisal can be used as a basis to exclude or down-weight flawed studies, and its outputs should be used in the synthesis in some way [5]: for example, by including study validity as a moderator or basis for sensitivity analysis in quantitative synthesis [e.g. 48], or in order to prioritise presentation and discussion of the evidence base. Complex scoring systems should be avoided to minimise the risk of introducing errors and to ensure repeatability. Instead, studies should be given categorical coding, for example *low*, *high* and *unclear* validity [49]. In addition, meta-analysis can be used to compare the magnitude of the effects in studies of different validity (e.g. observational and experimental studies). These analyses should not be used to adjust meta-analytical weighting but should inform judgements about overall strength of evidence and uncertainty in effect estimates.

7. *Inappropriate synthesis (using vote-counting and inappropriate statistics)*

Description: All literature reviews attempt to create new knowledge by summarising a body of evidence. For quantitative reviews this may take the form of a meta-analysis, i.e. combining of effect sizes and variances across all studies to generate one or more summary effect estimates with confidence intervals (or slopes and intercepts in the case of meta-regressions) [50]. Not all systematic reviews may use meta-analysis as a synthesis method, but all reviews that are identified as 'meta-analyses' must fulfil a number of standard requirements such as calculation of the effect sizes for individual studies, calculation of the combined effects and confidence intervals etc [51, 52]. Meta-analyses and systematic reviews are therefore overlapping, with some arguing that all meta-analyses in the environmental field should be based on systematic methods to identify, collate, extract information from and appraise studies as they are in other domains [53].

For reviews of qualitative evidence, summarising the body of evidence takes the form of a formal drawing together of qualitative study findings to generate hypotheses, create new theories or conceptual models [54]. The choice and design of the synthesis methods are just as critical to the rigour of a review as the question formulation, searching, screening, critical appraisal and data extraction: inappropriate synthesis invalidates all preceding steps. Where full synthesis is performed, authors should be careful to ensure they use established and appropriate synthesis methods.

One common problem with evidence syntheses occurs when authors fall foul of 'vote-counting' [reviewed in 55]. Vote-counting is the tallying-up of studies based on statistical significance and direction of their findings. This approach is problematic for several reasons. Firstly, it ignores statistical power and study precision. Many studies might report non-significant effect not because the effect does not exist, but because the statistical power of these studies is too low to detect it. Secondly, vote-counting ignores the magnitude of effect of each study: those showing a positive effect may have a much larger effect size than those showing a negative effect. Finally, vote-counting ignores study validity: the positive studies may have a much higher validity than the negative ones, for example due to better study designs.

Example: Sánchez-Bayo and Wyckhuys [20] claimed to have conducted a meta-analysis of studies on insect decline, but no standard meta-analysis methods were used and the review fails most criteria for meta-analyses [51, 52]. It is also unclear how annual decline rates were calculated, and such measures were not standard effect sizes. There is no mention of weighting, and ANOVA is inappropriate for combining estimates from different studies. Britt et al. [56] similarly did not use established meta-analysis methods in their quantitative synthesis.

Graham et al. [57] chose to use a vote-counting approach in their review on hedgerows as farmland habitats because "the data are too heterogeneous to allow any meaningful synthesis or meta-analysis… We follow a standard vote counting procedure where significant positive effects, significant negative effects, and no significant effects are assigned a 'vote' in order to integrate information and generalise the effect direction for each structural component on each taxonomic group". Delaquis et al. [58] similarly stated they deliberately chose a vote-counting approach, despite calculating effect sizes in some cases. Pacifici et al. [25] also synthesised by vote-counting to estimate the percentage of species in major groups that demonstrated responses to climate change. In their review of conservation intervention effectiveness, Burivalova et al. [32] visualised their mapping of evidence by displaying the number of studies for each intervention type and colour coding studies according to their direction of effect (positive, neutral, negative), thereby promoting so-called 'visual vote-counting'.

Mitigation strategies: Vote-counting should never be used instead of meta-analysis. If the data in primary studies are deemed to be too heterogenous to be combined by means of meta-analysis (e.g. because reported measures of outcome are too diverse), using a flawed approach such as vote-counting is unlikely to help. Instead, the scope of the review might need to be reassessed and narrowed down to a subset of studies that could be meaningfully combined. Alternatively, formal methods for narrative synthesis should be used to summarise and describe the evidence base [59]. It is perfectly acceptable (and encouraged) to tabulate the results of all studies in a *narrative synthesis* that includes quantitative results and statistical significance, but this should also include results of critical appraisal of study validity. Doing so ensures that no studies are 'excluded' from the review because data are not reported in a way that allows inclusion in a meta-analysis. Indeed, important conclusions can be made from narrative synthesis without meta-analyses [e.g. 60].

A common justification for vote-counting is lack of reporting of variance measures in ecological literature. Studies lacking variance measures should be included using the narrative synthesis methods described above. Where quantitative synthesis is desired, meta-analysis of a reduced dataset is preferable to vote-counting a larger data set, ignoring precision, effect magnitude and heterogeneity. Increasing provision of data as Open Science permeates ecological research practice should make this problem less pervasive in the future.

Maps of evidence (e.g. systematic maps) that aim to catalogue an evidence base typically do not extract study findings: this should primarily only be done in the context of a robust systematic review that also involves critical appraisal of study validity and, ideally, appropriate quantitative or qualitative synthesis. Only established qualitative and quantitative synthesis methods should be used making the most of the plethora of methodological support available in the literature.

8. A lack of consistency and error checking (working individually)

Description: An individual researcher performing the various tasks of a systematic review may interpret definitions, concepts and system boundaries differently from someone else. This variability is an inherent part of being human, but in a literature review it may result in the inclusion or exclusion of a different set of studies depending on individual interpretation. By working alone and unchallenged, a reviewer cannot be sure they are correctly interpreting the protocol. Similarly, working alone can lead to a higher rate of errors (and importantly for reviews, an unacceptable false negative error rate, or the erroneous exclusion of relevant studies) than working in concert with another researcher [61].

Example: In their review of the water chemistry habitat associations of the white-clawed crayfish (*Austropotamobius pallipes*), Rallo and García-Arberas [62] tabulated minima, maxima and mean for a range of water chemistry variables (their Table 4). Their review methods are not described, but there are several transcription errors in the table that should have been corrected by error checking or dual data extraction.

Mitigation: It is for the reasons of alternative interpretation and false negative errors that the major coordinating bodies require at least a subset of the evidence base to be processed (i.e. screening, data extraction and appraisal) by more than one reviewer – typically following by an initial trial of the task to ensure reviewers interpret and apply the instructions consistently (refining instructions where necessary to improve consistency) [5, 10]. Additionally, few individuals have the requisite skill set to acquire, appraise and synthesise studies alone. High quality evidence synthesis is likely to involve collaboration with information specialists, evidence synthesis methodologists/statisticians as well as domain specialists.

Advice for more rigorous reviews

In Box 1, we provide general advice for those involved in funding, commissioning, conducting, or editing/peer-reviewing/appraising a review. We give a number of specific recommendations to the research community to support rigorous evidence synthesis.

Box 1. Recommended actions for authors wishing to conduct more rigorous literature reviews.

- Familiarise yourself with the best practice in evidence synthesis methods and appreciate that systematic reviewing is a flexible methodology that can be applied to any research topic provided the question is suitably formulated.
- Make use of freely accessible guidance, minimum standards and educational resources provided by CEE and others (e.g. the Campbell Collaboration and Cochrane)
- Seek training in evidence synthesis to produce a reliable review with a lasting legacy and potential to impact decision-making
- Connect with existing communities of practice - individual methodologists, information specialists/librarians, working groups, specialist organisations, conferences - and make use of the plethora of online resources related to evidence synthesis
- Engage with stakeholders (including experts) when planning your review: consult with a broad range of stakeholders when setting the scope; with librarians and information specialists when developing the search strategy; with statisticians and synthesis methodologists when designing quantitative or qualitative synthesis; and with communications experts when translating review findings
- Ensure that a review is clear in its purpose and objectives

- Ensure the intended level of rigour (including transparency, procedural objectivity and comprehensiveness) of a review is achieved
- Follow Open Science principles when conducting and publishing reviews (Open Synthesis [63]) to ensure transparency, i.e. make your data, methods and paper freely accessible and reusable
- Check author guidance for specific journals for advice on what is requested to be included with systematic reviews, e.g. *Environmental Evidence*, which aims to publish high quality systematic reviews;
- Demonstrate and assess the rigour of a review and how it is reported using existing tools such as ROSES reporting standards [23], CEESAT (www.environmentalevidence.org/ceeder and CEE standards of conduct (http://www.environmentalevidence.org/information-for-authors)
- Editors and publishers should ensure that instructions for authors include sufficient detail and minimum standards regarding the conducting and reporting evidence syntheses, and they should ensure that authors follow them: for example, guidance for reviews for Biological Conservation state "Review articles… must include a methods section explaining how the literature for review was selected". Yet several recent reviews published in this journal lack methods section altogether [e.g. 26, 27]. Journals should endorse or enforce reporting and conduct standards, such as PRISMA (https://www.prisma-statement.org), ROSES (https://www.roses-reporting.com), or MECIR (https://methods.cochrane.org/methodological-expectations-cochrane-intervention-reviews)
- Methodology experts should support review authors and editors by: raising awareness of rigorous evidence synthesis methodology; developing and advertising Open Educational resources to support those wishing to conduct or appraise systematic reviews; acting as methodology editors and peer-reviewers for community journals (e.g. Environment International that has a dedicated systematic review editor); increasing efficiency of reporting and appraisal tools to make them easier to use in editorial triage and peer-review

---

Conclusions

Systematic reviews are increasingly seen as viable and important means of reliably summarising rapidly expanding bodies of scientific evidence to support decision making in policy and practice across disciplines. At the same time, however, there is a lack of awareness and appreciation of the methods needed to ensure systematic reviews are as free from bias and as reliable as possible, demonstrated by recent, flawed, high-profile reviews.

No one group is responsible for this failure and no one group produces perfect systematic reviews. We call for the entire research community to work together to raise the standard of systematic reviews published in conservation and environmental management. Whilst systematic reviews are significant undertakings that require careful planning and involvement of a range of experts, these are not reasons to abandon rigour in favour of an unregulated free-for-all in evidence synthesis methods. We call on review authors to conduct more rigorous reviews, on editors and peer-reviewers to gate keep more strictly, and the community of methodologists to better support the broader research community. We cannot afford to fund or generate second order research waste (i.e. poor-quality reviews): many primary studies are already a waste of resources [64], and we must not waste resources on methodologically poor or biased syntheses. Only by working together can we build and maintain a strong system of rigorous, evidence-informed decision-making in conservation and environmental management.

Competing interests

The authors declare they have no competing interests.

Acknowledgements

The authors thanks Chris Shortall from Rothamstead Research for useful discussions on the topic.

Author contributions

NRH developed the manuscript idea and a first draft. All authors contributed to examples and edited the text. All authors have read and approve of the final submission.

Figure legends

Figure 1. Schematic showing the mains stages necessary for the conduct of a systematic review as defined by the Collaboration for Environmental Evidence (www.environmentalevidence.org).

References

1. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal. 2009;26(2):91-108.
2. Haddaway NR, Macura B. The role of reporting standards in producing robust literature reviews. Nature Climate Change. 2018;8(6):444-7.
3. Pullin AS, Knight TM. Science informing Policy–a health warning for the environment. BioMed Central; 2012.
4. Haddaway N, Woodcock P, Macura B, Collins A. Making literature reviews more reliable through application of lessons from systematic reviews. Conservation Biology. 2015;29(6):1596-605.
5. Pullin A, Frampton G, Livoreil B, Petrokofsky G. Guidelines and standards for evidence synthesis in environmental management. Collaboration for Environmental Evidence. 2018.
6. White H. The twenty-first century experimenting society: the four waves of the evidence revolution. Palgrave Communications. 2019;5(1):47.
7. O'Leary BC, Kvist K, Bayliss HR, Derroire G, Healey JR, Hughes K, et al. The reliability of evidence review methodology in environmental science and conservation. Environmental Science & Policy. 2016;64:75-82.
8. Woodcock P, Pullin AS, Kaiser MJ. Evaluating and improving the reliability of evidence syntheses in conservation and environmental science: a methodology. Biological Conservation. 2014;176:54-62.
9. Collaboration C. Campbell systematic reviews: Policies and guidelines. Campbell Systematic Reviews. 2014;1.
10. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019.
11. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. bmj. 2017;358:j4008.
12. Haddaway NR, Land M, Macura B. A little learning is a dangerous thing": a call for better understanding of the term 'systematic review. Environment international. 2017;99:356-60.
13. Freeman RE. Strategic management: A stakeholder approach: Cambridge university press; 2010.

14. Haddaway NR, Kohl C, da Silva NR, Schiemann J, Spök A, Stewart R, et al. A framework for stakeholder engagement during systematic reviews and maps in environmental management. Environmental evidence. 2017;6(1):11.
15. Land M, Macura B, Bernes C, Johansson S. A five-step approach for stakeholder engagement in prioritisation and planning of environmental evidence syntheses. Environmental Evidence. 2017;6(1):25.
16. Oliver S, Dickson K. Policy-relevant systematic reviews to strengthen health systems: models and mechanisms to support their production. Evidence & Policy: A Journal of Research, Debate and Practice. 2016;12(2):235-59.
17. Savilaakso S, Garcia C, Garcia-Ulloa J, Ghazoul J, Groom M, Guariguata MR, et al. Systematic review of effects on biodiversity from oil palm production. Environmental Evidence. 2014;3(1):4.
18. Savilaakso S, Laumonier Y, Guariguata MR, Nasi R. Does production of oil palm, soybean, or jatropha change biodiversity and ecosystem functions in tropical forests. Environmental Evidence. 2013;2(1):17.
19. Haddaway NR, Crowe S. Experiences and lessons in stakeholder engagement in environmental evidence synthesis: a truly special series. BioMed Central; 2018.
20. Sánchez-Bayo F, Wyckhuys KA. Worldwide decline of the entomofauna: A review of its drivers. Biological conservation. 2019;232:8-27.
21. Agarwala M, Ginsberg JR. Untangling outcomes of de jure and de facto community-based management of natural resources. Conserv Biol. 2017;31(6):1232-46.
22. Gurevitch J, Curtis PS, Jones MH. Meta-analysis in ecology. 2001.
23. Haddaway NR, Macura B, Whaley P, Pullin AS. ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environmental Evidence. 2018;7(1):7.
24. Lwasa S, Mugagga F, Wahab B, Simon D, Connors JP, Griffith C. A meta-analysis of urban and peri-urban agriculture and forestry in mediating climate change. Current Opinion in Environmental Sustainability. 2015;13:68-73.
25. Pacifici M, Visconti P, Butchart SH, Watson JE, Cassola FM, Rondinini C. Species' traits influenced their response to recent climate change. Nature Climate Change. 2017;7(3):205-8.
26. Owen-Smith N. Ramifying effects of the risk of predation on African multi-predator, multi-prey large-mammal assemblages and the conservation implications. Biological Conservation. 2019;232:51-8.
27. Prugh LR, Sivy KJ, Mahoney PJ, Ganz TR, Ditmer MA, van de Kerk M, et al. Designing studies of predation risk for improved inference in carnivore-ungulate systems. Biological Conservation. 2019;232:194-207.
28. Li Y, Hu S, Chen J, Müller K, Li Y, Fu W, et al. Effects of biochar application in forest ecosystems on soil properties and greenhouse gas emissions: a review. Journal of Soils and Sediments. 2018;18(2):546-63.
29. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS med. 2009;6(7):e1000097.
30. Bernes C, Carpenter SR, Gårdmark A, Larsson P, Persson L, Skov C, et al. What is the influence of a reduction of planktivorous and benthivorous fish on water quality in temperate eutrophic lakes? A systematic review. Environmental Evidence. 2015;4(1):7.
31. McDonagh M, Peterson K, Raina P, Chang S, Shekelle P. Avoiding bias in selecting studies. Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]: Agency for Healthcare Research and Quality (US); 2013.

32. Burivalova Z, Hua F, Koh LP, Garcia C, Putz F. A critical comparison of conventional, certified, and community management of tropical forests for timber in terms of environmental, economic, and social variables. Conservation Letters. 2017;10(1):4-14.
33. Min-Venditti AA, Moore GW, Fleischman F. What policies improve forest cover? A systematic review of research from Mesoamerica. Global Environmental Change. 2017;47:21-7.
34. Bramer WM, Giustini D, Kramer BM. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. Systematic reviews. 2016;5(1):39.
35. Bramer WM, Giustini D, Kramer BM, Anderson P. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. Systematic reviews. 2013;2(1):115.
36. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research synthesis methods. 2020;11(2):181-217.
37. Livoreil B, Glanville J, Haddaway NR, Bayliss H, Bethel A, de Lachapelle FF, et al. Systematic searching for environmental evidence using multiple tools and sources. Environmental Evidence. 2017;6(1):1-14.
38. Mlinarić A, Horvat M, Šupak Smolčić V. Dealing with the positive publication bias: Why you should really publish your negative results. Biochemia medica: Biochemia medica. 2017;27(3):447-52.
39. Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics. 2018;74(3):785-94.
40. Haddaway NR, Bayliss HR. Shades of grey: two forms of grey literature important for reviews in conservation. Biological Conservation. 2015;191:827-9.
41. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of statistical software. 2010;36(3):1-48.
42. Bilotta GS, Milner AM, Boyd I. On the use of systematic reviews to inform environmental policies. Environmental Science & Policy. 2014;42:67-77.
43. Englund G, Sarnelle O, Cooper SD. The importance of data-selection criteria: meta-analyses of stream predation experiments. Ecology. 1999;80(4):1132-41.
44. Burivalova Z, Şekercioğlu ÇH, Koh LP. Thresholds of logging intensity to maintain tropical forest biodiversity. Current biology. 2014;24(16):1893-8.
45. Bicknell JE, Struebig MJ, Edwards DP, Davies ZG. Improved timber harvest techniques maintain biodiversity in tropical forests. Current Biology. 2014;24(23):R1119-R20.
46. Damette O, Delacote P. Unsustainable timber harvesting, deforestation and the role of certification. Ecological Economics. 2011;70(6):1211-9.
47. Blomley T, Pfliegner K, Isango J, Zahabu E, Ahrends A, Burgess N. Seeing the wood for the trees: an assessment of the impact of participatory forest management on forest condition in Tanzania. Oryx. 2008;42(3):380-91.
48. Haddaway NR, Hedlund K, Jackson LE, Kätterer T, Lugato E, Thomsen IK, et al. How does tillage intensity affect soil organic carbon? A systematic review. Environmental Evidence. 2017;6(1):30.
49. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. Bmj. 2011;343:d5928.
50. Stewart G. Meta-analysis in applied ecology. Biology letters. 2010;6(1):78-81.
51. Koricheva J, Gurevitch J. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology. 2014;102(4):828-44.

52. Vetter D, Ruecker G, Storch I. Meta-analysis: A need for well-defined usage in ecology and conservation biology. Ecosphere. 2013;4(6):1-24.
53. Stewart GB, Schmid CH. Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies. Research synthesis methods. 2015;6(2):109-10.
54. Macura B, Suškevičs M, Garside R, Hannes K, Rees R, Rodela R. Systematic reviews of qualitative evidence for environmental policy and management: an overview of different methodological options. Environmental evidence. 2019;8(1):1-11.
55. Koricheva J, Gurevitch J. Place of Meta-analysis among other Methods of research synthesis. In: Koricheva J, Gurevitch J, Mengersen K, editors. Handbook of Meta-analysis in Ecology and Evolution: Princeton Scholarship Online; 2013.
56. Britt M, Haworth SE, Johnson JB, Martchenko D, Shafer AB. The importance of non-academic coauthors in bridging the conservation genetics gap. Biological Conservation. 2018;218:118-23.
57. Graham L, Gaulton R, Gerard F, Staley JT. The influence of hedgerow structural condition on wildlife habitat provision in farmed landscapes. Biological Conservation. 2018;220:122-31.
58. Delaquis E, de Haan S, Wyckhuys KA. On-farm diversity offsets environmental pressures in tropical agro-ecosystems: a synthetic review for cassava-based systems. Agriculture, Ecosystems & Environment. 2018;251:226-35.
59. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme Version. 2006;1:b92.
60. Pullin AS, Bangpan M, Dalrymple S, Dickson K, Haddaway NR, Healey JR, et al. Human well-being impacts of terrestrial protected areas. Environmental Evidence. 2013;2(1):19.
61. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC medical research methodology. 2019;19(1):132.
62. Rallo A, García-Arberas L. Differences in abiotic water conditions between fluvial reaches and crayfish fauna in some northern rivers of the Iberian Peninsula. Aquatic Living Resources. 2002;15(2):119-28.
63. Haddaway NR. Open Synthesis: on the need for evidence synthesis to embrace Open Science. Environmental evidence. 2018;7(1):1-5.
64. Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. Bmj. 2018;363:k4645.

1. Question Formulation
2. Protocol (ideally, peer-reviewed and published)
3. Searching
4. Article Screening
5. Data Extraction
6. Critical Appraisal
7. Synthesis
8. Final Review (peer-reviewed and published)
9. Communication