The reliability of evidence review methodology in environmental science and conservation

Bethan C. O'Leary[a]*, Kristian Kvist[a], Helen R. Bayliss[a], Géraldine Derroire[b], John R. Healey[b], Kathryn Hughes[c], Fritz Kleinschroth[b], Marija Sciberras[c], Paul Woodcock[d], Andrew S. Pullin[a]

*corresponding author: bethancoleary@gmail.com

[a] Centre for Evidence-Based Conservation, Bangor University, LL57 2UW, UK

[b] School of Environment, Natural Resources and Geography, Bangor University, LL57 2UW, UK;

[c] School of Ocean Sciences, Bangor University, LL59 5AB, UK

[d] Joint Nature Conservation Committee, Monkstone House, Peterborough, PE1 1JY, UK

Abstract

Given the proliferation of primary research articles, the importance of reliable environmental evidence reviews for informing policy and management decisions is increasing. Although conducting reviews is an efficient method of synthesising the fragmented primary evidence base, reviews that are of poor methodological reliability have the potential to misinform by not accurately reflecting the available evidence base. To assess the current value of evidence reviews for decision-making we appraised a systematic sample of articles published in early 2015 (N=92) using the Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT). CEESAT assesses the methodology of policy-relevant evidence reviews according to elements important for objectivity, transparency and comprehensiveness. Overall, reviews performed poorly with a median score of 2.5/39 and a modal score of zero (range 0-30, mean 5.8), and low scores were ubiquitous across subject areas. In general, reviews that applied meta-analytical techniques achieved higher scores than narrative syntheses (median 18.3 and 2.0 respectively), as a result of the latter consistently failing to adequately report methodology or how conclusions were drawn. However, some narrative syntheses achieved high scores, illustrating that the reliability of reviews should be assessed on a case-by-case basis. Given the potential importance of reviews for informing management and policy, as well as research, it is vital that overall methodological reliability is improved. Although the increasing number of systematic reviews and meta-analyses highlight that some progress is being made, our findings suggest little or no improvement in the last decade. To motivate progress, we recommend that an annual assessment of the methodological reliability of evidence reviews be conducted. To better serve the environmental policy and management communities we identify a requirement for independent critical appraisal of review methodology thus enabling decision-makers to select reviews that are most likely to accurately reflect the evidence base.

Keywords

Evidence syntheses; Evidence-base; CEESAT; Review methodology; Decision-making; Review evaluation.

1. Introduction

Evidence reviews (defined in Table 1 and hereafter also referred to as reviews) in conservation and environmental science, as in other disciplines, are a vital tool to support decision making for researchers and decision-makers alike. Whereas more general literature reviews enable current states of knowledge to be summarised and trends and patterns across multiple datasets to be identified, evidence reviews focus on specific questions of the size and direction of effect achieved through an intervention or the impact of an action (whether desired or not). The value of evidence reviews to end-users is strongly dependent on review objectivity (i.e. the review methodology reduces the susceptibility of findings to bias, individual judgement, or prejudice) and comprehensiveness (Chalmers, 2003; Pullin and Knight, 2001; Rothstein et al., 2013). These qualities also assist researchers in identifying gaps in knowledge and areas of controversy or uncertainty, and can help decision-makers undertake informed management and defend potentially controversial or expensive actions (Gough et al., 2012). Where these qualities are not present, reviews have the potential to misinform and result in policies that have unwanted and unforeseen consequences and/or wasted research investment (Kirsch et al., 2008; Pullin and Knight, 2012), particularly if used as the single source of knowledge (although this will rarely be the case) or if selectively used by stakeholders with particular priorities. Avoiding such an eventuality imposes an obligation on those conducting evidence reviews to ensure their reliability and accurate reflection of the primary evidence base, and to transparently report review methodology to enable external assessment of reliability.

Evidence-based environmental policy is becoming a crucial element within wider societal debates on human impacts on the environment and future actions for environmental protection. Evidence may be used to inform policy from a number of sources including expert knowledge, experiential evidence, primary research, and review articles amongst others, each with their own potential biases and problems. For example, although expert knowledge may offer important guidance for non-specialists, experts can have biased opinions and their knowledge can lag behind published evidence (Ayyub, 2010). In addition, vested interests of multiple stakeholders can lead to selective use of evidence in political debates (e.g. Biber, 2012; Pielke, 2007; Sarewitz, 2004) giving an inflated impression of uncertainty of the science and reducing its potential to inform future policy. Similarly, while the decision-maker's own experiences or the experiences of others can provide valuable direction to decision-making, it may not be appropriate to generalise such experiences to different social, ecological or economic situations. Primary studies provide vital insight into the real-world application of, for example, a specific management intervention or conservation strategy under particular conditions, however increasing publication rates of primary literature (Larsen and von Ins, 2010; Li and Zhao, 2015; Pautasso, 2012) have resulted in ever-increasing evidence of variable quality for decision-makers to draw from. Effective and unbiased integration of published scientific evidence into policy and management is therefore impractical without evidence synthesis.

Based on our experience, we estimate that between 40 and 80 new review articles intending to inform decision-making were published each month (c. 480-960 per year) in the environmental peer-reviewed literature between 2012 and 2015. Multiple or overlapping reviews addressing the same basic issue or question are now commonplace (e.g. Claudet et al., 2008; Stewart et al., 2009) and misrepresentation of data within reviews resulting from conflicts of interest with funding organisations has been indicated (Wade et al., 2010). Perhaps more commonly, selection of primary

data to support an adopted position or belief (so-called 'policy-based evidence') may be consciously or subconsciously employed by review authors (Biber, 2012; Pullin and Knight, 2012). While the translation of evidence from science to policy is rarely linear and decisions are informed through other mechanisms as well as published literature (e.g. Sharman and Holmes, 2010; Wesselink et al., 2013 and references therein), misinformation and misrepresentation within reviews is likely to further undermine evidence-informed decision-making. There is consequently a need to develop ways in which the reliability of individual reviews can be evaluated and compared to determine the value of their contribution to the evidence base prior to their incorporation within the decision-making process. In addition, with so many reviews on environmental topics being published, it is valuable to have an overview of reliability that highlights both strengths in review conduct and opportunities for improvement.

The reliability of evidence reviews has been of concern in other sectors (e.g. Mulrow, 1987; Tranfield et al., 2003; vom Brocke et al., 2009) and, partly in recognition of this, systematic review methodology was developed in the health sector as a gold standard for collecting and synthesising evidence (Chalmers et al., 2002; Cook et al., 1997). This has subsequently been modified for other sectors (e.g. education and environment) to reflect the different methodological approaches employed. Systematic reviews follow strict guidelines (e.g. Collaboration for Environmental Evidence, 2013; The Cochrane Collaboration, 2011) designed to improve rigour and transparency, and to minimise biases to which more traditional reviews are susceptible. Subsequently a number of tools have been published for critically appraising and rating reviews against this best practice methodology (e.g. Guyatt et al., 2011; Shea et al., 2009; Woodcock et al., 2014). Within environmental science, most evaluations to date have focused on reviews within specific disciplines and that apply meta-analytical techniques (Huntington, 2011; Koricheva and Gurevitch, 2014; Philibert et al., 2012), identifying consistent weaknesses in conduct and reporting standards.

Based on environmental systematic review methodology, which is transferable to all reviews that use literature review techniques (Collaboration for Environmental Evidence, 2013), an assessment tool expressly intended for evaluating environmental evidence reviews has been developed (the Collaboration for Environmental Evidence Assessment Tool [CEESAT], Woodcock et al., 2014). CEESAT aims to evaluate review reliability by assessing methodological elements essential for objectivity, transparency and comprehensiveness to enable decision-makers to select reliable, unbiased reviews. Since systematic review methodology was introduced in the environmental sector a decade ago (Pullin and Stewart, 2006) its use has become more widespread (Haddaway et al., 2015). In this context, it is timely to take the opportunity to assess the current reliability of environmental evidence reviews.

We evaluated a snapshot of review articles published in early 2015 using CEESAT to examine the methodological reliability of environmental reviews. Our evaluation is restricted to a specific subset of the review literature that is intended to inform policy. For clarity, terminology related to evidence synthesis and reviews used in this article is defined in Table 1. We assess: (1) the reliability (objectivity, transparency, and comprehensiveness) of reviews (based on information reported within each review), and therefore the confidence that researchers and decision-makers (end-users) can place in the conclusions of these syntheses; (2) whether reliability varies according to the type of synthesis conducted or the structure of the question being answered; (3) whether there are differences in review reliability amongst broad subject areas (marine, terrestrial, freshwater); (4)

whether the ISI Journal Citation Reports impact factor can be used as a proxy by non-specialists for selecting more reliable reviews; and (5) the implications of our findings for end-users.

Table 1: Evidence synthesis and review terminology.

Term	Definition
Evidence review	An overarching term for articles that collate and summarise multiple primary studies related to a specific, policy-relevant question (Collaboration for Environmental Evidence, 2013).
Evidence synthesis	"A distinct element in the review process" that combines results from primary studies to derive findings from all available evidence. This "occurs once the evidence base has been accumulated and the data of interest extracted" (Pope et al., 2007, p15).
Meta-analysis	"A set of statistical methods for combining the magnitude of the outcomes (effect sizes) across different data sets addressing the same research question"(Koricheva et al., 2013, p8).
Narrative synthesis	A process which uses prose to summarise and draw conclusions from primary research and which may be supplemented by the reviewers' own experience. Some narrative syntheses may include limited quantitative analysis.
Systematic review	"A review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyse data from the studies that are included within the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies" (Collaboration for Environmental Evidence, 2013).

2. Methods

2.1 Selection of articles

A systematic literature search was undertaken in Web of Science, Scopus, CAB Direct and Google Scholar on 16th February and 16th March 2015. Searches were restricted to articles published in 2015 to obtain a sample of recent reviews. Only articles published in English were considered due to available resources. The search strategy is detailed in Table S.1.

All retrieved articles from Web of Science, Scopus and CAB Direct together with the first 100 hits from each Google Scholar search (following systematic review guidelines; Collaboration for Environmental Evidence, 2013) were screened for relevance. Titles and abstracts were screened according to the following inclusion criteria: (1) reviews should be undertaken in relation to a specific question or topic of relevance to environmental management and have recommendations for policy or practice; and (2) article type should be a review and/or synthesis of primary research.

Articles were first screened on their title to remove obviously irrelevant literature. Each retained article was then screened for relevance on the abstract. Recognising the potential for subjective decisions on article inclusion, a random sample of articles (20% of the total number to be screened on abstract) was also screened by a second person, with decisions compared using the kappa test of agreement (Cohen, 1960) to ensure repeatability of screening decisions. A kappa score of 0.76 was obtained, which indicates substantial agreement between reviewers and that decisions were sufficiently repeatable (Collaboration for Environmental Evidence, 2013). The few cases of disagreement were discussed to improve understanding of inclusion criteria prior to screening the remainder of articles.

2.2 Assessment of reviews

Relevant reviews were scored according to the Collaboration for Environmental Evidence Assessment Tool (CEESAT, Woodcock et al., 2014). CEESAT consists of a set of 13 criteria (Table 2) designed in alignment with environmental systematic review methodology (Collaboration for Environmental Evidence, 2013). Criteria aim to evaluate the reliability of reviews by assessing their objectivity, transparency and comprehensiveness. Note that by reliability we refer to the level of confidence an end-user may place in the review methodology rather than in the accuracy of review findings. Assessment of reviews was undertaken following the explanatory guidelines produced by Woodcock et al. (2014).

Reviews could receive 3, 1 or 0 points for each of the 13 criteria with a maximum of 39 points possible. All reviews (including appendices and supplementary materials) were scored by one scorer (referred to as scorer 1) and a second randomly selected scorer from a group of nine (scorers 2-10) to account for possible differences in the application of scoring criteria and potential biases introduced from scorer expertise. All scorers were environmental scientists ranging from Masters to Professorial level. Only two of the scorers (and not scorer 1) were involved in the design of CEESAT. In the case of disagreements between the two scorers about whether a criterion was met, the mean of their scores was used.

Reviews were classified as either a meta-analysis or narrative synthesis (which may contain some quantitative analysis; see Table 1) based on the type of synthesis conducted rather than the description provided within each review (see Koricheva and Gurevitch, 2014 for a discussion of uses and misuses of terminology). Reviews which apply meta-analytical techniques score 3 points in CEESAT criterion 6.1 (Table 2), narrative synthesis in which some quantitative analysis is conducted (e.g. graphical presentation or descriptive statistics) score 1 point while reviews which solely conduct narrative synthesis score 0 points. In addition, as many reviews are broad and may not be expected to provide a specific answer to a clearly defined question, lower scores may be explained by the lack of specific objective(s) in the review. A key component of systematic review methodology is the formulation of an appropriate answerable question. Questions for a systematic review are most commonly structured within a Population, Intervention/Exposure, Comparator, Outcome (PICO/PECO, henceforth referred to solely as PICO) closed-framed format (Collaboration for Environmental Evidence, 2013). All reviews were therefore also classified as to their question structure (PICO or open-framed) to consider whether clearly defined questions contributed to more reliable reviews.

Table 2: Summary of CEESAT criteria for scoring environmental reviews. For detailed rationale and guidelines for use see Woodcock et al. (2014).

	CEESAT criteria	Rationale
1	Was an *a-priori* protocol available for comment before the synthesis was conducted?	Prevents post hoc changes in methods, objectives and scope thereby increasing the robustness of review.
2: Searching for studies		
2.1	Does the search for literature utilise a comprehensive range of sources?	Increases the likelihood of capturing the available evidence base and reduces publication bias.
2.2	Are the search strings clearly defined?	Enables external evaluation, allows search to be repeated and avoids open-ended searches.
3: Including studies		
3.1	Does the synthesis apply clearly documented inclusion criteria to all potentially relevant studies found during the search?	Reduces the risk of subjective decisions regarding included studies. Reduces selection bias
3.2	Does the synthesis demonstrate that inclusion/exclusion decisions are repeatable?	Demonstrates the objectivity of study in/exclusion decisions.
3.3	Are inclusion/exclusion decisions transparent?	Enables external verification of in/exclusion decisions.
4: Critical Appraisal		
4.1	Does the synthesis conduct and report critical appraisals of the methods of each study?	Assesses the quality of evidence available for synthesis in terms of susceptibility to bias.
4.2	Are studies objectively weighted according to methodological quality?	Ensures that greater emphasis is given to more robust studies.
5: Data Extraction		
5.1	Is data extraction documented, repeatable and consistent?	Enables external validation and repetition and reduces the potential for bias.
5.2	Are the extracted data reported for each study?	Enables external verification and analysis of extracted data.
6: Data synthesis		
6.1	Is a quantitative synthesis conducted?	Increases objectivity by reducing the potential for subjective assessment of findings.
6.2	Is heterogeneity in the effect of the Intervention/Exposure investigated statistically?	Indicates the external/general applicability of results and appropriateness of combining studies.
6.3	Does the synthesis consider possible publication bias?	Assesses the potential for publication bias tc influence review findings.

2.3 Data analysis

The aim of this study is to provide an assessment of the methodological reliability of environmental reviews as a whole and to identify strengths and weaknesses in the population in order to offer guidelines for improving future reviews, not to praise or criticise individual reviews. Consequently, it

is important to note that we have not provided: (1) a list of included and excluded reviews; or (2) the individual scores achieved by each review.

Scoring decisions between scorer 1 and scorers 2-10 were compared using a weighted kappa test of agreement, an extension of the kappa analysis that takes into account the magnitude of disagreement between scorers (i.e. a 0-1 criterion score disagreement is given less weight in the equation [ranked as magnitude 1] than a 1-3 disagreement [magnitude 2] or a 0-3 disagreement [magnitude 3]). Final kappa values range between 0 and 1 and higher kappa values indicate greater agreement (Cohen, 1960; Landis and Koch, 1977).

We used descriptive statistics (median, mode and mean) to enable comparisons between synthesis type, subject area and question structure. Differences in the mean scores achieved by each review assigned to different categories (e.g. meta-analysis *vs.* narrative synthesis, subject area, etc.) were tested statistically. A Kruskal-Wallis test for multiple comparisons was used to examine relationships between subject areas and Mann-Whitney U tests were applied for pairwise comparisons (meta-analyses *vs.* narrative syntheses, PICO-structured *vs.* open-framed question). Pearson's correlation coefficient was used to analyse the relationship between 2014 ISI Journal Citation Reports impact factor and mean journal CEESAT score.

3. Results

3.1 Description of dataset

We identified 92 reviews meeting our inclusion criteria published between January and March 2015 across 68 different peer-reviewed journals and 3 grey literature sources. Reviews spanned terrestrial (58%), marine (14%) and freshwater (17%) realms with the remaining 11% (N=10) spanning multiple biomes or being non-specific. Narrative syntheses (85%, N=78) and meta-analyses (15%, N=14) were represented in the articles scored. Across all reviews 47% (N=43; comprising 72% narrative syntheses and 28% meta-analyses) used a PICO question structure. Meta-analyses were more likely to frame their review question in a PICO format (12/14) whereas 60% (N=47) of narrative syntheses used an open-framed format.

3.2 Review scores

Mean scores of individual reviews varied from 0 to 30 out of 39 (Fig. 1). Overall, the mean score was 5.8 but the median value was 2.5 and the modal value was zero.

Mean score achieved for each CEESAT criterion also varied, however overall scores were low with no criterion averaging greater than 1 (Fig.2a). Only criterion 3.1 (clearly documented inclusion criteria) achieved a median score greater than zero (0.5). All criteria achieved modal scores of zero.

3.3 Effect of synthesis type, question structure and subject area on review score

The median score of meta-analyses was 18.3 (range 8.5-30.0, mean 18.4) and for narrative syntheses was 2.0 (range 0.0-29.0, mean 3.5) with the difference being significant (Mann Whitney U=1067, N_1=78, N_2=14, $p<0.05$). Narrative syntheses including limited quantitative analysis (N=18) achieved higher median (6.8 *vs*. 1.0) and mean (8.1 *vs.* 2.2) scores than those with no quantitative analysis (N=60).

Meta-analyses achieved higher mean scores than narrative syntheses for all criteria except criterion 1 (availability of an *a-priori* protocol) (Fig. 2a). Median scores for each criterion encompassed the full range of possible scores (0-3) for meta-analyses. Narrative syntheses achieved a median score between zero and 0.25.

Reviews examining a PICO-structured question achieved higher median (5.5 *vs*. 2.5) and mean (8.4 *vs*. 3.5) scores than those with an open-framed question structure (Mann Whitney U=1371, N_1=49, N_2=43, $p<0.05$; Fig. 2b). Mean scores achieved by narrative syntheses using a PICO format were higher than open-framed across all criteria except 2.2 (clearly defined search strings) and 6.3 (publication bias) which were the same. Limited representation of meta-analyses addressing an open-framed question (2 open-framed *vs*.12 PICO) prevented comparison of the effect of question structure for this type of review.

The subject area did not have a major effect on the scores achieved by reviews with median and mean scores for: marine (5.0 and 7.1 respectively, N=13, range 0.5-19.5), freshwater (2.5 and 4.9 respectively, N=16, range 0.0-30.0), and terrestrial (2.5 and 6.0 respectively, N=53, range 0.0-24.0) (Kruskal Wallis H=2.37, 2 d.f., p=0.31). Mean scores of meta-analyses were higher than those of narrative syntheses in each subject area and PICO-structured syntheses achieved higher scores than open-framed reviews in terrestrial and freshwater subject areas (Fig. S.1).

3.4 Journal impact factor and review reliability

A slight, but significant positive correlation was found between the mean CEESAT score of individual reviews and the impact factor of the journal they were published in (Pearson's r=0.28, $p<0.05$, N=89). However, 10/14 meta-analyses were published in the 30 journals with the highest impact factor (5 in the top 20, 3 in the top 10) indicating that this relationship may be due to a greater proportion of reviews published in higher impact journals being meta-analyses. No significant correlation was found between journal impact factor and mean score when meta-analyses and narrative syntheses (Pearson's r=0.006, p=0.98, N=14 and r=0.13, p=0.25, N=78 respectively), or PICO or open-framed reviews (Pearson's r=0.29, p=0.056, N=43 and r=0.20, p=0.18, N=45 respectively) were tested separately.

3.5 Repeatability of scoring system

Scores assigned by different scorers were identified as being of 'moderate', 'substantial' or 'almost perfect' agreement (Landis and Koch, 1977) using a weighted kappa test of agreement indicating that scores were consistent between scorers (Table 3). Out of 1,196 assessments of individual criteria across reviews undertaken by nine scorers (scorers 2-10), only 15% (N=178) were subject to any disagreement with scorer 1. Of these, only 1% (N=14) were 0-3 point disagreements (i.e. one scorer considered a criterion to be fully met and assigned 3 points while the other did not and awarded 0 points). Scorer 1 was neither more lenient nor more strict in their assessments than scorers 2-10; scorer 1 awarded higher scores than scorers 2-10 in 48% (N=86) of scoring decisions where there was a disagreement (N=178) and lower scores in 52% (N=92) of disputed decisions (N=178). The mean absolute difference in total CEESAT scores awarded per article between scorers was only 1.3 indicating that even where differences were observed, the total score for each article was generally not substantially affected.

Table 3: Repeatability of scoring between scorer 1 and scorers 2-10, evaluated by kappa statistic weighted according to the extent of disagreement (e.g. a 0 *vs*. 1 or a 1 *vs*. 3 disagreement is less important than a 0 *vs*. 3 disagreement). Numbers represent the mean agreement across criteria for each combination of scorer 1 and the second scorer. An almost perfect agreement lies between 0.8 and 0.99, substantial between 0.6-0.8 and moderate between 0.4-0.6 (Landis and Koch, 1977). Reported to 2 d.p.

Scorer	2	3	4	5	6	7	8	9	10
Weighted kappa	0.64	0.43	0.72	0.65	0.47	0.75	0.68	0.81	0.52

4. Discussion

4.1 How reliable (objective, transparent, comprehensive) are reviews in the environmental sector?

Our results indicate that low CEESAT scores are ubiquitous in reviews published across the environmental sector and environmental journals. In accordance with previous appraisals (e.g. Roberts et al., 2006) we found that reviews do not consistently apply systematic methods to their review process or, if they do, their failure to report these methods adequately prevents high scores being achieved. Reviews scored highest for reporting of search strings (criterion 2.2) and the use of clearly defined inclusion criteria (criterion 3.1, Fig. 2). Conversely, all but one review in our sample failed to report an *a-priori* protocol (criterion 1), or fully demonstrate that their inclusion/exclusion decisions are repeatable (criterion 3.2) or transparent (criterion 3.3). Even for those criteria where the highest mean scores are achieved (criteria 2.2 and 3.1), only 42% of reviews provided sufficient detail to score any points.

4.2 Does the type of synthesis conducted affect review reliability?

Certain characteristics of reviews were found to lead to greater reliability. Meta-analyses achieve higher median and mean scores than narrative syntheses, as do reviews which applied a PICO question structure.

The higher scores achieved by meta-analyses are in part because certain criteria require statistical analysis to score highly (i.e. criteria 4: critical appraisal and 6: data synthesis) and because they are more likely to use a PICO question structure. However, points for these criteria are available to narrative syntheses and it is not uncommon for narrative syntheses to use a PICO question structure. Importantly, those narrative syntheses which did use a PICO question structure achieved higher mean scores than those that did not suggesting that this approach may aid the quality of reporting regardless of the type of synthesis. It is important to note, however, that some reviews in environmental science published as scientific papers may not have been primarily intended to be used directly in decision-making, and perhaps particularly reviews that use an open-framed question structure. Therefore, our inclusion criteria restricted the assessment to only those articles that made recommendations for policy or practice on the basis of literature review findings. By focusing on reviews whose findings are presented within a policy or practice decision-making context we ensured only those that are relevant for assessment using CEESAT were selected. The use of narrative syntheses within environmental policy and management has often been considered inappropriate due to their vulnerability to author bias and generally inadequate reporting of methodology (Bilotta et al., 2014; Lortie, 2014; Roberts et al., 2006). Such criticisms could be

addressed if narrative syntheses clearly reported their search strategies and documented extracted data. Narrative syntheses may be conducted if meta-analysis is not possible due to a lack of primary data and, if reliable, the former remain a useful component of the evidence base. For example, narrative syntheses such as those conducted under the Intergovernmental Panel on Climate Change and the Intergovernmental Platform on Biodiversity and Ecosystem Services are based on broad expert consultation which follows identified rules of conduct to improve credibility and help minimise identified shortcomings[1,2]. In addition, the highest scores were remarkably similar for both types of synthesis indicating that narrative syntheses are not doomed by default to low scores and that they can add to the reliable evidence base for researchers and decision-makers.

While scores were greater for meta-analyses and reviews which used a PICO question structure CEESAT scores remained low overall. The large variation in scores awarded to reviews suggest that neither the broad type of synthesis conducted nor the review question structure should be used as conclusive indicators for the reliability of a review.

4.3 Are marine, terrestrial or freshwater reviews more reliable?

Low scores were ubiquitous across subject areas with no major difference found in the reliability of reviews published in marine, terrestrial or freshwater disciplines. However in general, the type of synthesis conducted and the review's question structure did influence review reliability in each discipline according to the same trend as noted previously (Fig. S.1). This implies that no one broad research community is publishing more reliable reviews than another, and instead this variation is more attributable to individual author decisions about applied methodology or reporting. Review methodological standards require improvement across the environmental sector.

4.4 Can a journal's impact factor be used as a proxy to select more reliable reviews?

The impact factor of a journal is a widely used indicator for evaluating scientific journals (Zupanc, 2014) and could therefore be seen as a way to identify more reliable reviews for decision-making. Impact factors attract heavy criticism (e.g. Zupanc, 2014) and decision-makers have been advised to exercise caution in basing research evaluation on the journal within which a review is published (Guerrero, 2001; Jarwal et al., 2009). Nevertheless, we considered the relationship between journal impact factor and review reliability in order to inform non-specialists who might consider using journal impact factor as a proxy for selecting more reliable reviews. While we found that journals with a higher impact factor do publish reviews which scored significantly higher, the magnitude of this effect was small and this relationship did not continue when assessed by review synthesis type. Moreover, journal policies differ on the format and guidelines for review articles which may have influenced this result in a way that does not reflect differences in the underlying methodology used in the review process. The effect may also be an artefact of journals with a higher impact factor publishing a greater proportion of meta-analyses (relative to narrative syntheses), although further data would be required to test this potential source of bias rigorously. In addition, we noted variation in review reliability within each journal where we could compare more than two articles.

[1] IPCC Principles and Procedures. Available at: http://www.ipcc.ch/organization/organization_procedures.shtml [accessed 22/03/2016].
[2] IPBES Guidance and Conceptual Framework. Available at: http://www.ipbes.net/guidance-and-conceptual-framework [accessed 22/03/2016].

These results indicate that the reliability of environmental reviews should not be based on the impact factor of the journal they are published in.

4.5 Limitations

The repeatability of assessments (Table 3) between scorers imparts confidence in review reliability scores generated using CEESAT, as well as suggesting that external users (albeit with a science background) could apply CEESAT themselves when considering the value of a review. It is, however, important to note that CEESAT is a relatively crude measure and does not take into account some key aspects of reliability such as use of appropriate statistical techniques, methodological or interpretation errors, or fraud, as well as other properties that may influence the value to non-specialists (e.g. clarity of writing). Nevertheless, we feel that CEESAT's emphasis on susceptibility to bias is appropriate and it has performed adequately to answer the questions posed. See Woodcock et al. (2014) for a detailed discussion of caveats and considerations when interpreting CEESAT scores.

5. Conclusions

Evidence reviews can be valuable for informing future research direction and decisions at all levels of environmental management (Cook et al., 2012; Pullin et al., 2004; Seavy and Howell, 2010) and, while reviews are not always used directly in policy formation, they can be influential in the development of the evidence base and indirectly influence the policy debate (e.g. Dicks et al., 2014; Land et al., 2016). However, our sample of reviews suggests that many published evidence reviews are of low methodological reliability which increases the risk that they will not adequately reflect current knowledge. Such reviews thus have the potential to misinform decision-making, especially if selectively used by stakeholders with particular priorities. We found that important information describing methodology and results was frequently missing, with narrative syntheses performing particularly poorly overall. The consistent lack of transparency and methodological rigour of evidence reviews reduces the confidence end-users should place in the findings of reviews published within the field of environmental science in general.

Our results show that certain methodological characteristics can lead to greater review reliability although as CEESAT scores remained low overall, and there was large variation in scores awarded, none of these characteristics should be used as conclusive indicators for review reliability. Consequently, when selecting reviews for decision-making reliability should be considered on a case-by-case basis, preferably using a standard critical appraisal tool such as CEESAT. While decisions are often made with restricted timescales and resources, on average it took our team 30 minutes to complete one critical appraisal with CEESAT, suggesting that this could be incorporated into the decision-making process.

There are multiple possible reasons why otherwise rigorously conducted reviews may fail to adequately report methodology (and therefore be assessed as less reliable). Publication restrictions (e.g. word limits) often apply, particularly in higher impact journals, however greater use could be made of supplementary information to provide a full description of review methodology (e.g. covering all the CEESAT scoring criteria; Table 2). In addition, we found evidence that reporting requirements are perceived differently for different types of synthesis; dedicated methodology sections were only included in 36% of narrative syntheses compared with all meta-analyses

assessed. However, there is no reason why descriptions of methods cannot also be routinely incorporated into narrative syntheses. Moreover, reporting standards for evidence reviews are published in other sectors such as medicine (Moher et al., 2009). An absence of these in the environmental sector might mean that expectations on the type of information to include may be lower and more variable. Nonetheless, sufficient lessons can be learned from other disciplines, systematic reviews in environmental science, and from CEESAT to improve standards.

Decision-making is informed by a number of factors including other types of evidence than review articles, which are introduced at various stages of the decision-making process, and trade-offs amongst them. Consequently, while improving access to more reliable reviews will improve the evidence available to decision-makers, decisions will still be informed by several mechanisms, and improved decision-making may not necessarily occur. The relevance or applicability of a review may also sometimes be given greater weight by decision-makers over a review's methodological reliability. In such circumstances, it is important that decision-makers are aware of any weaknesses in review conduct to determine how much weight should be given to relevance over methodological reliability. Furthermore, where a decision-maker is faced with multiple relevant reviews we suggest that greater weight should be placed on the methodologically more robust reviews. Our experience of working with organisations and individuals directly involved in environmental decision-making shows that evidence reviews contribute substantially to policy and management decisions. Indeed, organisations such as the UK Government Department for Environment, Food and Rural Affairs often commission evidence reviews on particular questions to inform their decision-making (e.g. Buller et al., 2015; Newman et al., 2015). Reviews that are directly commissioned to inform policy should undoubtedly be conducted in a way that ensures methodological rigour. However, reviews that relate to decision-making more broadly should also aim to accurately represent the primary evidence base and report their methodology as clearly as possible. If decision-making is to be informed by evidence, it is important to ensure that this evidence is a good representation of the whole evidence base (Dicks et al., 2014).

Given the introduction of more rigorous review methodology and reporting standards into the environmental sector a decade ago (Pullin and Stewart, 2006; Roberts et al., 2006) the current variation in review reliability is perhaps disappointing. Whilst there are examples of reviews that receive moderate-high scores with CEESAT, the number of reviews scoring poorly should be of concern to environmental research and policy stakeholders. Ways by which authors of reviews can improve the reliability of their syntheses mainly relate to better reporting and have been detailed extensively elsewhere (e.g. Haddaway et al., 2015; Koricheva and Gurevitch, 2014; Philibert et al., 2012; Roberts et al., 2006). These improvements include aspects such as the development of an *a-priori* protocol, reporting of search strategies and inclusion criteria, and detailing the included studies and extracted data. To motivate progress in improving review reliability we advocate an annual assessment of published reviews using this study as a 2015 baseline. Given the proliferation of publication outlets for reviews it is unlikely that low-reliability reviews will be eliminated from the literature and consequently the use of peer-reviewed publication as the sole indicator of reliability will continue to be flawed. One option for addressing this problem for the environmental policy and management communities could be to establish an independent database of evidence reviews, which includes an assessment of reliability for each review based on CEESAT or a similar system. Such a database could be open access and would facilitate access to the most up-to-date and reliable evidence.

As the review literature expands, assessments of reliability are likely to become increasingly important and topical, and given the importance that may be accorded to reviews and their potential to misrepresent the evidence base, it is vital that we improve this key mechanism for making science available to policy.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We would like to thank the editor and an anonymous reviewer for their insightful comments that greatly improved this manuscript.

References

Ayyub, B.M., 2010. On uncertainty in information and ignorance in knowledge. International Journal of General Systems 39, 415-435.

Biber, E., 2012. Which science? Whose science? How scientific disciplines can shape environmental law. The University of Chicago Law Review 79, 471-552.

Bilotta, G.S., Milner, A.M., Boyd, I., 2014. On the use of systematic reviews to inform environmental policies. Environmental Science & Policy 42, 67-77.

Buller, H., Hinchliffe, S., Hockenhull, J., Barrett, D., Reyher, K., Butterworth, A., Heath, C., 2015. Systematic review and social research to further understanding of current practice in the context of using antimicrobials in livestock farming and to inform appropriate interventions to reduce antimicrobial resistance within the livestock sector. Department for Environment, Food & Rural Affairs, Report O00558.

Chalmers, I., 2003. Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up to date evaluations. Annals of the American Academy of Political and Social Sciences 589, 22-40.

Chalmers, I., Hedges, L.V., Cooper, H., 2002. A brief history of research synthesis. Evaluation & The Health Professions 25, 12-37.

Claudet, J., Osenberg, C.W., Benedetti-Cecchi, L., Domenici, P., Garcia-Charton, J.A., Perez-Ruzafa, A., Badalamenti, F., Bayle-Sempere, J., Brito, A., Bulleri, F., Culioli, J.-M., Dimech, M., Falcon, J.M., Guala, I., Milazzo, M., Sanchez-Meca, J., Somerfield, P.J., Stobart, B., Vandeperre, F., Valle, C., Planes, S., 2008. Marine reserves: size and age do matter. Ecology Letters 11, 481-489.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

Collaboration for Environmental Evidence, 2013. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. Environmental Evidence.

Cook, C.N., Carter, R.W., Fuller, R.A., Hockings, M., 2012. Managers consider multiple lines of evidence important for biodiversity management decisions. Journal of Environmental Management 113, 341-346.

Cook, D.J., Mulrow, C.D., Haynes, R.B., 1997. Systematic reviews: synthesis of best evidence for clinical decisions. Annals of Internal Medicine 126, 376-380.

Dicks, L.V., Walsh, J.C., Sutherland, W.J., 2014. Organising evidence for environmental management decisions: a '4S' hierarchy. Trends in Ecology & Evolution 29, 607-613.

Gough, D., Oliver, S., Thomas, J., 2012. An introduction to systematic reviews. Sage, London.

Guerrero, R., 2001. Misuse and abuse of journal impact factors. European Science Editing 27, 58-59.

Guyatt, G.H., Oxman, A.D., Schunermann, H.J., Tugwell, P., Knottnerus, A., 2011. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. Journal of Clinical Epidemiology 64, 380-382.

Haddaway, N.R., Woodcock, P., Macura, B., Collins, A., 2015. Making literature reviews more reliable through application of lessons from systematic reviews. Conservation Biology epub.

Huntington, B.E., 2011. Confronting publication bias in marine reserve meta-analyses. Frontiers in Ecology and the Environment 9, 375-376.

Jarwal, S.D., Brion, A.M., King, M.L., 2009. Measuring research quality using the journal impact factor, citations and 'Ranked Journals': blunt Instruments or Inspired metrics? . Journal of Higher Education Policy and Management 31, 289-300.

Kirsch, I., Deacon, B.J., Huedo-Medina, T.B., Scoboria, A., Moore, T.J., Johnson, B.T., 2008. Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. PLoS Med 5, e45.

Koricheva, J., Gurevitch, J., 2014. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology 102, 828-844.

Koricheva, J., Gurevitch, J., Mengersen, K., 2013. Handbook of meta-analysis in ecology and evolution. Princeton University Press.

Land, M., Granéli, W., Grimvall, A., Hoffmann, C.C., Mitsch, W.J., Tonderski , K.S., Verhoeven, J.T.A., 2016. How effective are created or restored freshwater wetlands for nitrigen and phosphorus removal? A systematic review. Environmental Evidence 5, 9.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159-174.
Larsen, P.O., von Ins, M., 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84, 575-603.
Li, W., Zhao, Y., 2015. Bibliometric analysis of global environmental assessment research in a 20-year period. Environmental Impact Assessment Review 50, 158-166.
Lortie, C.J., 2014. Formalized synthesis opportunities for ecology: systematic reviews and meta-analyses. Oikos 123, 897-902.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6, e1000097.
Mulrow, C.D., 1987. The medical review article: state of the science. Annual International Medicine 106, 485-488.
Newman, J.R., Duenas-Lopez, M.A., Acreman, M.C., Palmer-Felgate, E.J., Verhoeven, J.T.A., Scholz, M., Maltby, E., 2015. Do on-farm natural, restored, managed and constructed wetlands mitigate agricultural pollution in Great Britain and Ireland? A systematic review. Department for Environment, Food & Rural Affairs, Report WT0989.
Pautasso, M., 2012. Publication growth in biological sub-fields: patterns, predictability and sustainability. Sustainability 4, 3234-3247.
Philibert, A., Loyce, C., Makowski, D., 2012. Assessment of the quality of meta-analysis in agronomy. Agriculture, Ecosystems and Environment 148, 72-82.
Pielke, R.A., 2007. The honest broker: Making sense of science in policy and politics. Cambridge University Press.
Pope, C., Mays, N., Popay, J., 2007. Synthesizing qualitative and quantitative health evidence. A guide to methods. McGlaw Hill.
Pullin, A.S., Knight, A.T., Stone, D.A., Charman, K., 2004. Do conservation managers use scientific evidence to support their decision-making? Biological Conservation 119, 245-252.
Pullin, A.S., Knight, T.M., 2001. Effectiveness in conservation practice: pointers from medicine and public health. Conservation Biology 15, 50-54.
Pullin, A.S., Knight, T.M., 2012. Science informing policy - a health warning for the environment. Environmental Evidence 1, 15.
Pullin, A.S., Stewart, G.B., 2006. Guidelines for systematic review in conservation and environmental management. Conservation Biology 20, 1647-1655.
Roberts, P.D., Stewart, G.B., Pullin, A.S., 2006. Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. Biological Conservation 132, 409-423.
Rothstein, H.R., Lortie, C.J., Stewart, G.B., Koricheva, J., Gurevitch, J., 2013. Quality standards for research syntheses, In: Koricheva, J., Gurevitch, J., Mengersen, K. (Eds.), Handbook of meta-analyses in ecology and evolution. Princeton University Press. Princeton.
Sarewitz, D., 2004. How science makes environmental controversies worse. Environmental Science & Policy 7, 385-403.
Seavy, N.E., Howell, C.A., 2010. How can we improve information delivery to support conservation and restoration decisions? Biodiversity and Conservation 19, 1261-1267.
Sharman, A., Holmes, J., 2010. Evidence-based policy or policy-based evidence gathering? Biofuels, the EU and the 10% target. Environmental Policy and Governance 20, 309-321.
Shea, B.J., Hamel, C., Wells, G.A., Bouter, L.M., Kristjansson, E., Grimshaw, J., Henry, D.A., Boers, M., 2009. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. Journal of Clinical Epidemiology 62, 1013-1020.

Stewart, G.B., Kaiser, M.J., Cote, I.M., Halpern, B.S., Lester, S.E., Bayliss, H.R., Pullin, A.S., 2009. Temperate marine reserves: global ecological effects and guidelines for future networks. Conservation Letters 2, 243-253.
The Cochrane Collaboration, 2011. Cochrane handbook for systematic reviews of interventions, In: Higgins, J.P.T., Green, S. (Eds.). Version 5.1.0.
Tranfield, D., Denyer, D., Smart, P., 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management 14, 207-222.
vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., Cleven, A., 2009. Reconstructing the giant on the importance of rigour in documenting the literature search process, ECIS 2009 Proceedings. Paper 161.
Wade, L., Whitehead, H., Weilgart, L., 2010. Conflict of interest in research on anthropogenic noise and marine mammals: does funding bias conclusions? Marine Policy 34, 320-327.
Wesselink, A., Buchanan, K.S., Georgiadou, Y., Turnhout, E., 2013. Technical knowledge, discursive spaces and politics at the science-policy interface. Environmental Science & Policy 30, 1-9.
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.
Zupanc, G.K.H., 2014. Impact beyond the impact factor. Journal of Comparative Physiology-A 200, 113-116.

Figure Captions

Fig. 1: Distribution of reviews by overall CEESAT score awarded, subdivided by synthesis type; narrative syntheses (white) and meta-analyses (black). Mean score presented in 0.5 intervals to the maximum available score of 39.

Fig. 2: Mean scores for (a) all syntheses (grey), meta-analyses (black) and narrative syntheses (white), and (b) PICO-structured (grey) and open-framed (white) reviews across CEESAT criteria.