Biological Conservation 176 (2014) 54–62

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon

Evaluating and improving the reliability of evidence syntheses in conservation and environmental science: A methodology

Paul Woodcock [a], Andrew S. Pullin [a,*], Michel J. Kaiser [b]

[a] Centre for Evidence-Based Conservation, School of Environment, Natural Resources and Geography, Bangor University, Gwynedd LL57 2UW, UK
[b] School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey LL59 5AB, UK

ARTICLE INFO

Article history:
Received 13 September 2013
Received in revised form 31 March 2014
Accepted 23 April 2014
Available online 6 June 2014

Keywords:
Evidence base
Environmental policy
Meta-analysis
Evidence synthesis
Systematic review
Review methodology

ABSTRACT

The volume of primary literature in conservation and environmental science is expanding rapidly. Evidence syntheses that review and combine the findings from research on policy-relevant questions are therefore vital for informing decision-making. However, such syntheses exhibit considerable variation in conduct and reporting, potentially undermining their value to decision-makers. To address this problem, we developed a scoring system – the Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT) – that uses detailed criteria and guidelines to evaluate policy-relevant syntheses in conservation and environmental science. The higher the score awarded, the greater the objectivity, comprehensiveness and transparency of the synthesis, hence the greater the confidence in its reliability. We then used 40 review articles to test CEESAT in terms of (i) applicability to different syntheses, (ii) validity of scores awarded, (iii) effectiveness at discriminating between syntheses, and (iv) repeatability of scoring by different assessors. CEESAT was applicable to 36 articles, and scores ranged from 1 to 33 (mean = 13.2, median = 15, maximum possible = 39). Variation in overall scores and in the individual criteria shows that CEESAT discriminates effectively among syntheses, making differences in rigour clear. Scoring was repeatable, indicating that assessments are not overly susceptible to differences in the application and interpretation of guidelines. The detailed rationale and guidelines for each criterion should help improve future syntheses and promote consistent scoring between assessors. Furthermore, scores can be used directly by non-specialists to compare syntheses that investigate key conservation questions and to incorporate reliability and rigour into decision-making.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The volume of conservation and environmental research available to decision-makers is expanding rapidly, and has reached a point where focusing on selected primary studies is unlikely to provide a reliable reflection of all of the available evidence (Fazey et al., 2004; Lawler et al., 2006 and Trimble and van Aarde, 2012). Evidence syntheses that review and combine the findings from primary research articles to assess the effectiveness of an environmental intervention (e.g. habitat restoration) or the impact of an exposure (e.g. pollution) are therefore increasingly important for consolidating research. Such syntheses may be narrative/qualitative reviews or quantitative reviews (e.g. meta-analyses), but differ from other types of review article in which the principal focus is on using existing research to develop new concepts, and from technical reports that are not expressly conducted as syntheses of primary research. Accordingly, policy-relevant evidence syntheses represent a vital link between researchers and decision-makers – particularly when the results of individual studies appear inconsistent (Pullin and Knight, 2009 and Sutherland et al., 2004). However, the practical value of this link hinges on the extent to which syntheses provide an objective and comprehensive reflection of all available research (Pullin and Knight, 2001).

Importantly, there is considerable variation in the conduct of evidence syntheses in conservation and environmental science (Gates, 2002; Huntington, 2011; Philibert et al., 2012 and Roberts et al., 2006). This variation reflects differences in how studies are searched for, how information in these studies is extracted and synthesised, and how the methods used to produce the synthesis are reported. Such differences affect the rigour of evidence syntheses in terms of objectivity, comprehensiveness and transparency, and consequently can lead to incomplete or inaccurate conclusions (Englund et al., 1999; Lajeunesse and Forbes, 2003; Stewart, 2010 and Whittaker, 2010). If policymakers are unable to conduct in-depth critical evaluation of each relevant synthesis, the most

* Corresponding author. Tel.: +44 1248 382444.
E-mail address: a.s.pullin@bangor.ac.uk (A.S. Pullin).

http://dx.doi.org/10.1016/j.biocon.2014.04.020
0006-3207/© 2014 Elsevier Ltd. All rights reserved.

rigorous may be overlooked and weaknesses underestimated, potentially leading to inappropriate management and policy decisions (Pullin and Knight, 2012). Access to evaluations of syntheses in conservation and environmental science are therefore likely to be valuable to decision-makers and other non-specialists, provided that such evaluations have been conducted transparently and to consistent guidelines (e.g. by trained assessors). Standardised guidelines could also be used independently by researchers to objectively identify questions lacking current high quality syntheses, or by editors or peer-reviewers for preliminary assessment of manuscripts.

Systematic review methodology, which comprises a series of guidelines expressly designed to ensure high levels of objectivity and transparency in review methods is regarded as a 'gold standard' for conducting evidence syntheses (Cochrane Collaboration, 2011; CEE, 2013 and Moher et al., 2009) and therefore provides an appropriate basis for assessing reliability. Indeed, the utility of modified systematic review guidelines in evaluating evidence syntheses is increasingly recognised in other fields, particularly medicine (e.g. Burda et al., 2011; Popovich et al., 2012; Shea et al., 2007 and Zumsteg et al., 2012). These evaluation tools are expressly designed for assessing systematic reviews, and, as such, employ stringent scoring criteria with relatively brief guidance on their application. In conservation and environmental science, however, the majority of reviews are non-systematic and exhibit a diverse range of methods, objectives and approaches to reporting. Consequently, it is important that assessment systems include criteria that identify differences amongst non-systematic reviews (as well as discriminating amongst systematic reviews), and that detailed guidance on applying criteria is provided in order to ensure their consistent and objective use. In addition, differences between fields mean that systematic review guidelines have required adaptation before their application in conservation (Fazey et al., 2004). Scoring systems based on guidelines developed expressly for conducting systematic reviews in conservation (e.g. CEE, 2013) and tested and refined using a range of relevant syntheses are thus more likely to be appropriate for evaluating reviews in the field than existing approaches that are based on guidelines from other fields (e.g. the AMSTAR tool developed in medicine (Popovich et al., 2012 and Shea et al., 2007)).

Whilst evaluations of evidence syntheses have been conducted in conservation and environmental science, these were targeted at researchers and concentrated on identifying strengths and weaknesses in methodology (e.g. Huntington, 2011; Philibert et al., 2012 and Roberts et al., 2006). As such, the objective was principally to guide the development of future reviews, rather than to provide a tool explicitly intended for repeated use in review assessments. A detailed explanation and rationale for each criterion used is therefore often lacking, making it difficult to apply these systems objectively and consistently. As one example, both Roberts et al. (2006) and Philibert et al. (2012) use subjective terms in defining what is acceptable in the search strategy for a review – the former requests that a 'detailed systematic literature search of sources (e.g. use of specific databases and/or reference lists)' is documented (criterion 6) and the latter requests 'correct description of the bibliographic search procedures used by the authors to select the individual studies (i.e. papers) and the repeatability of these procedures (criterion 1)'.

Here, we adapt systematic review guidelines and incorporate recommendations from previous studies (Liberati et al., 2009; Roberts et al., 2006 and Shea et al., 2007) to propose 13 criteria relevant to the evaluation of the objectivity, transparency and comprehensiveness of policy-relevant evidence syntheses in conservation and environmental science. These criteria have been developed, tested, and refined using reviews specific to the field and we intend that they are applied to evidence syntheses that seek to assess the effectiveness of an environmental intervention or the impact of an exposure on a non-target. We emphasise that the criteria are designed to assess the utility of syntheses for decision-making, and not their overall scientific merit (which would also require the incorporation of features such as conceptual insights and methodological advances). Our criteria also differ from systematic review guidelines and existing scoring systems (e.g. Liberati et al., 2009; Roberts et al., 2006; Shea et al., 2007 and Zumsteg et al., 2012) in that we provide detailed explanatory guidelines that are explicitly intended to allow objective and repeatable scoring of the conduct of both systematic and non-systematic reviews in conservation and environmental science. For each criterion, we therefore describe optimal, intermediate and inadequate standards, and we provide rationale and examples. Collectively, we term the scoring criteria, guidelines, and examples, the Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT).

2. Materials and methods

2.1. Proposed scoring guidelines for the Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT)

Guidelines for applying each of the 13 scoring criteria are given in Table 1 (below). Syntheses receive 3 points (GREEN), 1 point (AMBER), or 0 points (RED) for each criterion. Rationale for each criterion is provided in Text A.1, together with examples. The following definitions derived from systematic review methodology are used: Population – 'The taxa, community, ecosystem, process or property under study'. Intervention/Exposure – 'An action or agent with possible impact on a Population. It may be potentially negative (e.g. pollution) or potentially positive (e.g. habitat restoration)'. Outcome – 'The measures used to quantify how a Population is affected by an Intervention/Exposure'.

2.2. Selection of reviews and test of scoring system

To determine the viability of the Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT), we first selected 40 reviews in conservation and environmental science (see Table A.1 for bibliography). We did not consider reviews that predominantly examined theories/concepts rather than policy questions, articles clearly marked as opinion pieces, or reports/policy guidelines that did not explicitly synthesise primary research. Because this exercise is intended primarily to test CEESAT, rather than to evaluate the quality of all reviews on a particular question, articles were deliberately chosen (based on the title and abstract) to encompass both qualitative and quantitative syntheses in a range of policy-relevant fields. Although patterns in the scores awarded may highlight areas for further consideration, the results do not, therefore, indicate the general strengths and weaknesses of reviews in their respective fields. All reviews were scored by PW and a subset of 20 were independently scored by a second assessor. ASP was involved in the design of CEESAT and was the second assessor for 10 reviews. For contrast, MJK was not involved in the design of CEESAT, received no instruction other than the CEESAT guidelines, and was the second assessor for the remaining 10 reviews. We tested CEESAT with respect to four elements:

(i) *Applicability:* The scoring system should be applicable to any policy-relevant evidence synthesis that investigates whether or not an environmental intervention or anthropogenic impact has a particular effect. Applicability was examined by considering whether each criterion could be clearly assessed in all reviews, and whether the review addressed

specific questions of potentially high direct value to policymakers.

(ii) *Validity:* The total score of a review should reflect its' objectivity and transparency. By using the systematic review framework, CEESAT is based on an externally recognised approach to promoting objectivity and transparency. Nonetheless, there are many ways in which systematic review guidelines could be applied. Consequently, we followed Shea et al. (2009) and assessed CEESAT by comparing review scores with those obtained using related tools. In the absence of explicitly developed review evaluation systems for conservation and environmental science, a single assessor (PW) also scored all reviews using the 27-point Yes/No checklist in Roberts et al. (2006), hereafter 'ROB' (with the caveat that ROB does not provide detailed explanations for how to apply each criterion). We then tested for a correlation between the overall scores obtained using each approach (for ROB, Yes = 1, No = 0).

(iii) *Effectiveness:* The scoring system should transparently discriminate between reviews to make any differences in rigour apparent.

(iv) *Repeatability:* For a given review, assessors should obtain similar or identical scores for individual criteria and in total.

2.3. Data analysis

With the exception of the test of repeatability, scores and analyses are based on assessments by PW. Spearman's rank correlations were conducted to test for a correlation between CEESAT scores and those awarded using ROB. Effective discrimination was investigated by visually inspecting the distribution of total scores and calculating descriptive statistics, and by examining the scores for individual criteria. If a criterion only ever obtains the same score across all test reviews then it may have limited discriminatory power and be deemed redundant – particularly if consistently scoring 3 points. To investigate repeatability in scoring, we used a Spearman's rank correlation to test for consistency in the total score awarded by PW and the second assessors. We also investigated repeatability between assessors in the scoring of individual criteria by comparing agreement in the scores awarded using (i) % agreement, (ii) kappa analysis (maximum kappa score is 1, and high kappa scores indicate high agreement between assessors, Cohen, 1960 and Landis and Koch, 1977) and (iii) weighted kappa analysis (which takes into account the magnitude of disagreement between assessors – e.g. an AMBER-RED disagreement is ranked as magnitude 1, and a GREEN–RED disagreement as magnitude 3; Shea et al., 2007 and Viera and Garrett, 2005).

3. Results

3.1. Summary of reviews

The reviews selected covered both negative anthropogenic impacts (e.g. rainforest degradation) and conservation interventions (e.g. wetland restoration) and both quantitative and qualitative syntheses were represented (Table A.1). Reviews were generally well-cited and published in journals with an impact factor $\geqslant 2$. These reviews are therefore likely to be perceived as rigorous and to be in areas of interest to policymakers.

3.2. Applicability of scoring system

On full-text examination, 36 of the 40 reviews were determined to be policy-relevant evidence syntheses suitable for evaluation. The remaining reviews were judged to have a largely conceptual emphasis and/or to provide broad outlines of research areas rather than focusing on specific policy questions and so we did not regard CEESAT as applicable. Note that some of the evidence syntheses that were scored also examined questions additional to those stated in Table A.1, and in some instances these questions were not appropriate for all elements of the scoring system. Scores therefore reflect only aspects of syntheses that address the questions stated in Table A.1.

3.3. Validity of scoring system

Using CEESAT, overall scores of policy-relevant evidence syntheses ranged from 1–33 out of 39 (mean = 13.2 ± 1.5, median = 15; Fig. 1). Using alternative scoring in which GREEN was awarded 2 or 4 points resulted in narrower and wider distributions of scores, but did not alter rankings and so we continued with a 3–1–0 system for subsequent analyses (see Table A.1). The total scores for individual syntheses were correlated with those obtained using ROB (Spearman's rho = 0.94, $p < 0.001$), supporting the use of CEESAT as a valid mechanism for assessing syntheses.

3.4. Effectiveness of scoring system

The wide range of scores indicates that CEESAT has discriminated effectively between syntheses. Whilst this discrimination partly reflects CEESAT distinguishing between meta-analyses and narrative syntheses, there is also marked variation within each of these two categories of synthesis, including an outlying narrative synthesis with a higher score than several meta-analyses (Peppin et al., 2011, Fig. 1).

There was considerably greater variation in CEESAT scores than in those obtained using ROB, and this pattern held across all syntheses (CEESAT $\sigma = 9.2$, ROB $\sigma = 4.8$; Fig. 1) as well as within meta-analyses (CEESAT $\sigma = 5.2$, ROB $\sigma = 2.1$) and narrative syntheses (CEESAT $\sigma = 4.7$, ROB, $\sigma = 2.9$). Each individual criterion in CEESAT contributed to discriminating between syntheses, although for several criteria maximum points were rarely awarded (Fig. 2).

3.5. Repeatability of scoring system

The total score awarded to each synthesis by PW and ASP, and by PW and MJK were closely correlated (PW and ASP: Spearman's rho = 0.91, $p < 0.001$; PW and MJK: Spearman's rho = 0.95, $p < 0.001$; Fig. 3). PW scored slightly higher than the second assessors but the mean difference in scores awarded was low (2.2 ± 0.4). Agreement was also high for most of the individual criteria (100% agreement for three criteria and at least 70% for a further six criteria; Table 2). Although unweighted kappa scores for some criteria were low, these values increased when kappa was weighted to take into account the extent of disagreement between assessors.

4. Discussion

Evidence syntheses that combine the findings from primary research on the effects of interventions (e.g. habitat restoration) or exposures (e.g. pollution) are increasingly recognised as vital to informing conservation decisions (Pullin and Knight, 2009 and Sutherland et al., 2004). However, these syntheses employ a diverse range of approaches (reporting and conducting) and vary widely in rigour. The Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT) assesses evidence syntheses using criteria adapted from the CEE guidelines for conducting rigorous systematic reviews in conservation and environmental

Table 1
Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT) criteria and scoring guidelines. For each criterion, reviews score 3 points (GREEN), 1 point (AMBER) or 0 points (RED) according to how well the criterion is met. See Appendix Text A.1 for further details.

1 Protocol	A protocol is a document produced prior to the commencement of an evidence synthesis. It describes the background to the synthesis, the questions, the strategy that will be used to search for primary research articles, and the criteria for deciding whether or not an article is then relevant to include in the synthesis. The protocol should also outline the approach to assessing the quality of each included study, and to extracting and synthesising data from primary research articles (CEE, 2013). Writing a protocol is therefore analogous with developing and documenting a methodology prior to conducting fieldwork or experiments and is similarly integral to producing a study that is robust against *post hoc* changes in methods and scope (CRD, 2009 and Liberati et al., 2009).
1.1 *Was an a-priori protocol available for comment before the synthesis was conducted?*	
GREEN (3)	An *a-priori* protocol is linked from the synthesis (e.g. as supplementary material or online).
AMBER (1)	N/A
RED (0)	No *a-priori* protocol is available.
2 Searching for studies	An optimal search for literature should possess three key properties: comprehensive (maximises the number of potentially relevant studies found), systematic (avoiding *ad hoc* search strategies reduces the susceptibility to bias resulting from e.g. no defined endpoint of search) and transparent (readers should be able to repeat and evaluate the search).
2.1 Does the search for literature utilise a comprehensive range of sources?	
GREEN (3)	Documents use of resources capturing both peer-reviewed and grey literature. - Peer-reviewed literature: At least three databases or two databases and systematic bibliography searches.- Grey literature: Systematically searches relevant websites or uses specified internet search engine(s). Statements such as '*We considered only peer-reviewed material because this is more reliable than grey literature*' without evidence that the methodological quality of potentially relevant grey literature was assessed do not indicate that grey literature was objectively considered.
AMBER (1)	Documents use of resources capturing peer-reviewed literature. Should use at least two relevant databases OR one database and systematic search of websites or bibliographies.
RED (0)	Uses a single database without bibliography search or does not document the use of databases.
2.2 Are the search strings clearly defined?	
GREEN (3)	All search terms, Boolean operators ('AND', 'OR' etc.) and wildcards clearly stated so that the exact search is repeatable by a third party.
AMBER (1)	Clear evidence of a search, but the search is only partially repeatable by a third party either because *(i)* specific search terms are not stated or *(ii)* Boolean operators/wildcards are not stated (so it is unclear how the search terms are combined).
RED (0)	Search is vaguely defined or undefined. Repeatability is low or not possible. Describing the background or objectives of the synthesis does not indicate that a search has taken place.
3 Including studies	Comprehensive searches may generate a large number of articles that vary widely in their relevance to the synthesis. Authors must then determine whether or not each article is sufficiently relevant for inclusion in the data synthesis stage. However, the choice of inclusion criteria can influence the conclusions of the synthesis, and the application of inadequately defined criteria can be subjective (Englund et al., 1999; Lajeunesse and Forbes, 2003 and Whittaker, 2010). Decisions over which studies are relevant for inclusion should therefore be based on clearly defined criteria, and should be repeatable and transparent. Criteria 3.1–3.3 refer only to studies included/excluded on the basis of relevance – see point 4.2 for inclusion/exclusion on the basis of methodological quality.
3.1 Does the synthesis apply clearly documented inclusion criteria to all potentially relevant studies found during the search?	
GREEN (3)	Clear that *a priori* criteria for filtering the articles found during the search are systematically applied to all potentially relevant articles. Criteria should be precisely defined (e.g. reliance on broad and potentially ambiguous terms such as 'ecosystem functioning' should be avoided).
AMBER (1)	The questions/scope/objectives for the synthesis are stated such that the type of primary research articles to be included are broadly apparent, but the synthesis does not explicitly identify *a priori* inclusion criteria to be systematically applied to all articles found during the search.
RED (0)	It is not clear from the introduction and objectives which primary research articles should be included/excluded.
3.2 Does the synthesis demonstrate that inclusion/exclusion decisions are repeatable?	
GREEN (3)	Inclusion/exclusion criteria are independently applied by more than one person to some or all of the studies located during the search. Kappa statistic (CEE, 2013; Cohen, 1960 and Landis and Koch, 1977) or related metric is calculated and indicates good repeatability.
AMBER (1)	As above, but kappa statistic or related metric indicates a low-moderate degree of repeatability OR inclusion decisions carried out by more than one person but results of repeatability test not presented.
RED (0)	Repeatability not tested.
3.3 Are inclusion/exclusion decisions transparent?	
GREEN (3)	Lists all studies found during the search and explains the decision for excluded studies. This information should be provided for all studies that were read at full-text but subsequently excluded from the synthesis.
AMBER (1)	Lists all studies included in the synthesis AND lists some (at least one) of the individual studies that were excluded, together with explanations for the exclusions. Alternatively, lists all included and excluded studies, but does not explain the reasons for exclusion.
RED (0)	Does not list the studies included in the synthesis OR does not explain exclusion decision for any individual study.
4 Critical appraisal	Primary research can vary widely in methodological quality. This variation can influence the findings of the research, and, if not properly accounted for, the conclusions of syntheses that use it (Gates, 2002 and Lajeunesse, 2010). Critical appraisal involves transparently evaluating the design of each study, and weighting of studies based on methodologies can then help to objectively account for variation in study quality by placing greater emphasis on the most reliable studies (Pullin and Knight, 2003 and Norris et al., 2012).
4.1 Does the synthesis conduct and report critical appraisals of the methods of each study?	
GREEN (3)	Objectively and transparently evaluates the rigour of all relevant studies using pre-defined criteria (e.g. Pullin and Knight, 2003). The criteria will vary according to the synthesis question but critical appraisal should result in an explicitly documented assessment of study quality that incorporates the internal or external validity of each included study. Internal validity might consider sampling effort (e.g. study duration, number of replicates) and study design (e.g. collection of pre- and post-Intervention data from multiple sites). External validity might consider how generalisable the findings from an article are, e.g. spatial scale and distribution of study sites in relation to the synthesis question. This does not cover syntheses in which the methods for each study are stated but validity is not explicitly considered, or in which methodological rigour is discussed without transparent and objective assessments for each study.
AMBER (1)	Documents relevant information on methodology but does not explicitly and systematically assess internal or external validity for each study. Information relevant to validity should be provided (e.g. study design, sampling effort, spatial scale, study region, taxa). Information on methodological techniques (e.g. census methods, equipment used) is only relevant if the synthesis indicates that the choice of technique can influence study validity.
RED (0)	Does not document information on study design or sampling effort for all studies.

(*continued on next page*)

4.2 Are studies objectively weighted according to methodological quality?

GREEN (3)	Defined and repeatable approach to objectively accounting for differences in study quality: Weighting: e.g. In meta-analyses using inverse variance, sample size, rigour of study design etc. The metric used to weight studies should be clearly stated. Design: e.g. use aspects of study design/sampling effort as predictors of effect size, analyse groups of studies separately according to critical appraisal outcomes or conduct sensitivity analyses with and without methodologically weaker studies. Where methodology (study design, sampling effort) is incorporated into weighting, or where different study designs are treated separately, the details on which this treatment is based should be provided for each individual study.
AMBER (1)	Studies are treated differently according to methodological differences:Weighting: Weighting based on methodology, but weights are not stated for each study (e.g. weights based on sample sizes but sample sizes not reported) so weighting is not fully transparent.Study Removal: The only approach to weighting is the removal of methodologically flawed studies before synthesis. Removals should be clearly explained and based on methodological quality (sample size etc.), not methodological relevance (did not generate the metrics required by the synthesis etc.). The latter are covered by 3.1–3.3.Discussing differences in methodologies between studies is not equivalent to objective, quantitative weighting.
RED (0)	No evidence that methodological quality of primary research articles has been objectively incorporated into data synthesis. Includes syntheses that justify a focus on published research as the sole mechanism of ensuring article quality, as well as commentaries on the methods of individual studies that do not lead to a quantitative weighting.

5 Data Extraction	The volume and type of data collected by primary research articles varies substantially, even when similar questions are addressed. Authors of evidence syntheses must make decisions on which results to extract and on how to extract this information. These decisions may influence the findings of the synthesis (Gates, 2002 and Whittaker, 2010), and so to minimise bias the approach to data extraction should be clearly stated and, wherever possible, the extracted metrics should be comparable and consistent between studies.

5.1 Is data extraction documented, repeatable and consistent?

GREEN (3)	The methods (procedures and rules) by which metric(s) were extracted from each included study are stated. To ensure that data extraction is repeatable and objective, the synthesis should clearly indicate an intention to systematically extract a set of defined metrics from each research article.
AMBER (1)	The synthesis does not provide a fully repeatable *a priori* methodology for systematic data extraction, but the metrics extracted from each study can be determined (e.g. a table that lists all studies synthesised and states the Outcome metric for each study might be included).
RED (0)	Does not indicate an intention to extract particular metrics, and the metrics used are inconsistent or unclear. The synthesis might summarise the findings of several studies without presenting data.

5.2 Are the extracted data reported for each study?

GREEN (3)	A table of extracted data sufficient to inform/explain any subsequent narrative or quantitative synthesis is provided. States quantitatively the selected outcome metrics (or the effect size), and the Population and Intervention for each study.
AMBER (1)	A table that includes some of the extracted metrics for some or all studies is provided. At least two of the Outcome (can be qualitative), Population or Intervention are stated.
RED (0)	Synthesis does not provide information on at least two from the Outcome, Population and Intervention as specified above. Includes syntheses in which some information on the Population/Intervention/Outcome is given only in the text in a non-systematic manner.

6 Data Synthesis	The approach to synthesising included studies varies substantially, and some approaches are more effective at ensuring objectivity and minimising potential bias than others.

6.1 Is a quantitative synthesis conducted?

GREEN (3)	The effects of the Intervention in each individual study are quantitatively synthesised and statistically compared through meta-analysis or equivalent techniques (e.g. Dent and Wright 2009; Halpern, 2003 and Maliao et al., 2009).
AMBER (1)	The effects of the Intervention in each individual study are quantitatively synthesised (e.g. graphically or using other descriptive statistics) but not statistically compared OR quantitative synthesis is considered but determined to be inappropriate/not possible. Restating the results from each piece of primary research does not constitute a quantitative synthesis.
RED (0)	Data synthesis is exclusively qualitative. Also covers syntheses that test whether or not an Intervention has an effect by comparing the number of significant and non-significant studies – this 'vote-counting' approach has limited value because it focuses only on p-values and does not take into account the magnitude of the effect in each study.

6.2 Is heterogeneity in the effect of the Intervention/Exposure investigated statistically?

GREEN (3)	The effects of variables other than the Intervention/Exposure (e.g. taxa being considered, location, habitat type, study design etc.) are investigated statistically. Alternatively, no evidence for heterogeneity between studies is found (e.g. following calculation of Q statistic).
AMBER (1)	N/A
RED (0)	Effect modifiers are not statistically assessed. Includes syntheses that provide qualitative assessments of the importance of effect modifiers.

6.3 Does the synthesis consider possible publication bias?

GREEN (3)	Uses an objective statistical test to assess the likelihood of publication bias in the existing literature (e.g. Egger test) and evaluates the robustness of the synthesis conclusions to potential publication bias (e.g. fail-safe number of non-significant studies needed to alter the conclusions) (Rosenberg, 2005 and Borenstein et al., 2009).
AMBER (1)	Synthesis either *(i)* assesses the likelihood of publication bias statistically (e.g. Egger test) or *(ii)* assesses the potential effect of publication bias (e.g. calculates failsafe numbers). Also includes syntheses that investigate the likelihood of publication bias subjectively (e.g. constructs funnel plot) and syntheses that systematically contact authors of original articles for raw datasets (see Text A.1 for rationale on the latter).
RED (0)	Does not address publication bias using any of the broad classes of approach available. Statements such as '*We considered only peer-reviewed material because this is more reliable than unpublished studies*' without evidence that the methodological quality of potentially relevant unpublished studies was assessed do not indicate that grey literature was objectively considered.

science (CEE, 2013). To ensure that CEESAT provides meaningful and valid evaluations of systematic and non-systematic reviews in the field, scoring criteria were tested and refined extensively using syntheses on a range of relevant topics, and detailed guidance on the application of each criterion was developed. CEESAT discriminated effectively among policy-relevant syntheses and provided a clear indication of reliability that was consistent between assessors. We evaluate CEESAT in detail below, and examine how scores could be interpreted and utilised.

4.1. Evaluation of CEESAT

4.1.1. Applicability

CEESAT was tested on reviews with widely differing objectives, methodologies and transparency of reporting, and from diverse fields. Some potentially relevant reviews were not assessed because they explored several topics and often included detailed ecological theory and conceptual discussions (Table A.1). There may be ambiguity with respect to determining whether or not it

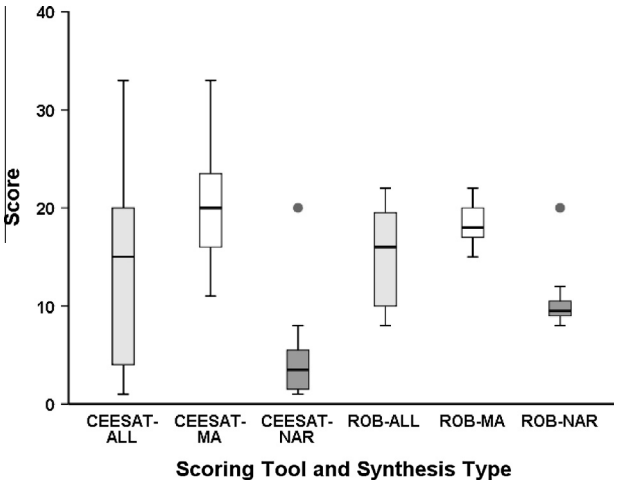

Fig. 1. Box and whisker plots showing distribution of scores awarded to evidence syntheses using CEESAT (max = 39) and Roberts et al. (2006) (ROB, max = 27). Median, interquartile range and range (excluding outliers) are shown. 'ALL': All syntheses, 'MA': Syntheses with meta-analyses, 'NAR': Syntheses without meta-analysis. Narrative syntheses contained one statistical outlier (Peppin et al., 2011), indicated by a grey circle.

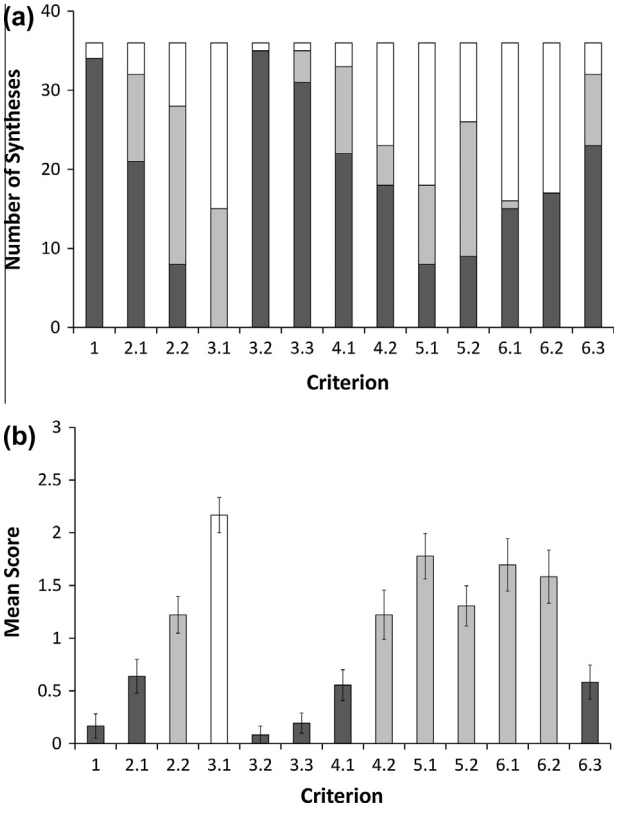

Fig. 2. Scores awarded for each criterion in CEESAT: (a) number of syntheses scoring 3 points (white), 1 point (light grey) or 0 points (dark grey) and (b) mean scores ± S.E. for each criterion. Scores are from assessments by PW only.

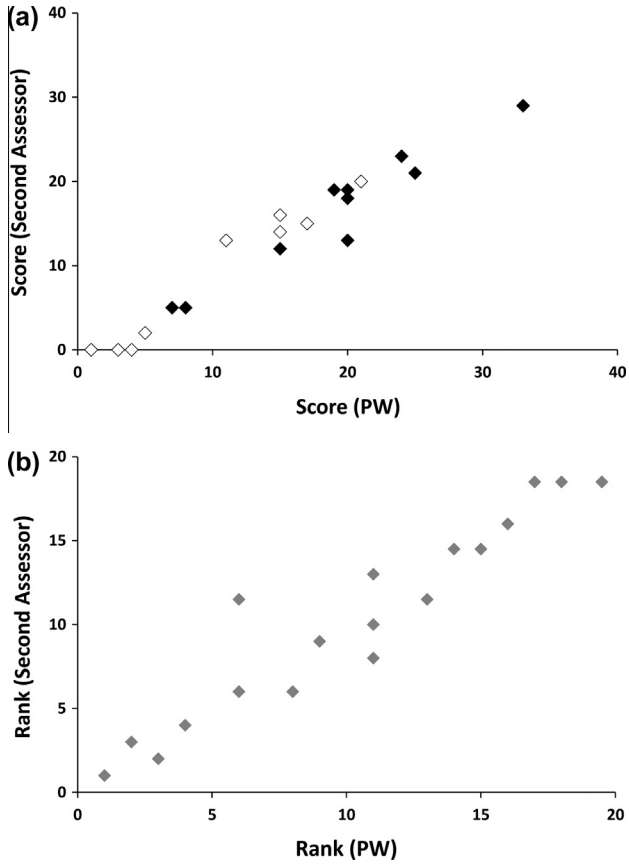

Fig. 3. Total scores and ranks for each synthesis as awarded by PW and by the second assessors: (a) total scores awarded. Second assessor is ASP (filled symbols) and MJK (open symbols) and (b) ranks for each synthesis based on scores awarded by PW and by the second assessors.

Table 2
Repeatability of scoring for individual criteria, evaluated by % agreement between assessors, by unweighted kappa statistic, and by kappa statistic weighted according to the extent of disagreement (e.g. a 0 vs 1 disagreement is less important than a 0 vs 3 disagreement). See Section 2 and Text A.1 for detailed explanations of criteria.

Criteria	% Agreement	Kappa	Weighted Kappa
1.1 Protocol	100	1	1
2.1 Search resources	95	0.92	0.91
2.2 Search string stated	65	0.49	0.67
3.1 Documented inclusion criteria	65	0.44	0.73
3.2 Evidence that inclusion decisions repeatable	100	1	1
3.3 Documented exclusion decisions	80	0.38	0.38
4.1 Critical appraisal of methods	75	0.55	0.61
4.2 Objective weighting	85	0.75	0.83
5.1 Data extraction documented	60	0.37	0.61
5.2 Extracted data reported	55	0.34	0.41
6.1 Quantitative synthesis	100	1	1
6.2 Heterogeneity investigated	90	0.81	0.83
6.3 Publication bias considered	70	0.21	0.46

is appropriate to apply CEESAT to a particular review, and so we recommend three approaches for assessors in this position: (i) based on reading the full-text of the review, two assessors should independently decide whether the review scope encompasses sufficient relevant material to address specific questions of potentially high direct value to policymakers, with disagreements resolved by a third party or by consultation with the review authors, (ii) a conservative approach in which 'uncertain' reviews are noted but not scored is employed and (iii) reviews excluded before scoring are recorded and the reasons for exclusion noted. Users should also consider the appropriateness of CEESAT with respect to their own rationale for evaluating the review: if the primary interest is not objectivity, transparency and comprehensiveness, then CEESAT will be less appropriate.

4.1.2. Validity and effectiveness

Despite using less than half the number of criteria, CEESAT was more effective at discriminating between syntheses than the checklist in Roberts et al. (2006) (Fig. 1) – probably in part because the latter was not expressly intended as a scoring tool. Although the two systems produced broadly consistent results, we thus regard the detailed criteria and guidelines in CEESAT as a more appropriate methodology for the evaluation of policy-relevant evidence syntheses. Whilst discrimination using CEESAT partly reflected whether or not a meta-analysis was conducted, substantial variation amongst meta-analyses and amongst narrative syntheses demonstrates that scores do not solely reflect the methodology used. This also illustrates that meta-analysis does not guarantee a rigorous conclusion if other steps in the review process are inappropriate (Harrison, 2011). Similarly, the relatively high score for Peppin et al. (2011) indicates that even where meta-analysis is not possible, many other aspects of reviews can be conducted robustly and are recognised as such by CEESAT.

Each individual criterion contributed to discriminating between syntheses, and for most criteria a spread of scores was achieved (Fig. 2). However, three criteria all averaged <0.5. These criteria may therefore represent key areas where standards of conduct and reporting can be improved (see Text A.1 for rationale). Furthermore, criteria 1 and 3.3 ask syntheses to formally and transparently provide information that is often developed informally. Some degree of pre-planning will take place prior to the commencement of an evidence synthesis (criterion 1), and authors are likely to make decisions on whether or not potentially relevant studies should be excluded (3.3) (Harrison, 2011). If this information is available and is robust then the credibility of syntheses is enhanced by presenting it in full (as an online Appendix if necessary). Likewise, it is relatively straightforward to test the repeatability of decisions over which of the studies located during the search are relevant for inclusion in the synthesis (CEE, 2013 and Landis and Koch, 1977). Note also that similarly informative techniques are available both to assess the possibility of publication bias and its' potential impact on review findings (Rosenberg, 2005 and Borenstein et al., 2009). The rigour of evidence syntheses can therefore be improved without being prohibitively time-consuming if criteria 3.2 and 6.3 are routinely met. More generally, the CEESAT scoring criteria can be used as a complement to the detailed systematic review guidelines (CEE 2013) to assist authors and editors in ensuring that future reviews are as objective and comprehensive as possible.

4.1.3. Repeatability of scoring

Scoring systems developed to assess the reliability of systematic reviews in medicine document high repeatability between assessors for total scores and for most criteria, but lower levels of agreement for some individual criteria (e.g. Popovich et al., 2012; Shea et al., 2007 and Shea et al., 2009). Similarly, total scores for individual syntheses using CEESAT were closely correlated between assessors and there was high agreement for most criteria (Table 1 and Fig. 3). Importantly, this suggests that assessments are not overly susceptible to differences in the interpretation or application of guidelines. Higher repeatability scores using weighted kappa than with unweighted kappa (Table 1) also indicates that disagreement tends to be with respect to whether RED (0 points) or AMBER (1 point) should be awarded – such disagreements therefore have a limited influence on total scores. As with other scoring systems however, some criteria were more problematic to score consistently. This emphasises the importance of using at least two independent assessors where possible, with a robust mechanism for resolving any disagreements. The scoring guidelines are also likely to evolve as any consistently problematic criteria are identified and clarified. For example, wording in the AMSTAR tool developed by Shea et al. (2007) to evaluate systematic reviews in medicine was updated to improve the repeatability of some criteria in response to comments from users (Shea et al., 2009) and a further modified version (r-AMSTAR) subsequently tested (Popovich et al., 2012).

4.2. Interpreting and using CEESAT scores

Total scores provide summary information that can be used to gauge the reliability of a synthesis that is being used to inform conservation policy. However, our intention is that CEESAT scores are used primarily as a comparative tool, to assist decision-makers in selecting the synthesis that is most likely to provide an accurate reflection of current research on key conservation policy questions. As an example of the comparative approach, the use of umbrella species as a tool for protecting less charismatic species is an important part of many conservation plans (Veríssimo et al., 2011 and references therein). In testing CEESAT, we included two syntheses examining the effectiveness of umbrella species: Branton and Richardson (2011) and Roberge and Angelstam (2004). These reviews were published in the same journal, and the latter was more heavily cited in 2013 (17 times vs 6 times: Web of Knowledge 15 January 2014). However, Branton and Richardson (2011) scored considerably higher (Appendix Table A.1), indicating that this is the more rigorously conducted synthesis and therefore should be focused on by decision-makers.

In addition to the total score, considering the scores for individual criteria can be informative. The way in which these 'profiles' are interpreted depends on the objectives and priorities of the user. We view each of the six sections in Table 2 as being of approximately equal importance, and so would favour syntheses that score moderately well in each section over those with high scores in some sections and low scores in others: for example, a meta-analysis is of limited value if it is based on a subjectively selected or incomplete portion of the literature. Users of CEESAT will also need to interpret scores in the context of their specific interest – a relatively low scoring synthesis may still be valuable if the subject matter is more directly relevant to the user.

4.3. Caveats

CEESAT evaluates evidence syntheses with respect to objectivity, transparency and comprehensiveness, and so cannot identify issues such as analytical errors or misconduct, or other aspects of syntheses (e.g. conceptual advances) which are particularly important in furthering research but may have less direct relevance to policymakers. In some instances, low scores may also reflect limitations in the primary literature. For example, inadequate reporting of methodological details and study variances in primary research restricts options for critical appraisal. Low-scoring syntheses can thus still contain valuable insights, perspectives and background, and may be important in driving more focused syntheses or additional primary research. We also emphasise that only two of the syntheses assessed here were explicitly conducted to full CEE systematic review guidelines (Kalies et al., 2010 and Peppin et al., 2011), and consequently that only these syntheses were expressly intended to meet guidelines similar to the CEESAT criteria. It is therefore encouraging that several other syntheses still achieved relatively high scores, and our hope is that openly available CEESAT scores will assist policymakers in focusing on such articles.

CEESAT does not discriminate between syntheses that do not meet a criterion and syntheses that may meet a criterion but do not report it. However, we feel that many of the scoring criteria would generally be reported if met (particularly for Green): it is unlikely, for example, that a review would conduct a meta-analysis

but not report doing so. Furthermore, inadequate reporting represents a lack of transparency. Transparency is key to the integrity of the review process, and so confidence in the reliability of inadequately reported reviews is diminished relative to reviews in which key information is clearly reported. To promote improved conduct of future reviews, we argue that transparent reporting of relevant information should be appropriately recognised compared with instances in which this information is unclear. Nonetheless, we do recognise that there may be occasions where relevant information has been omitted by review authors, particularly if unaware of the CEESAT criteria. The CEESAT evaluations used by policymakers would therefore be conducted transparently by registered assessors, with review authors given the opportunity to challenge the scores awarded prior to finalisation of the assessment, by directing assessors to information that may have been overlooked. Lastly, it is important to note that scores reflect likelihoods, rather than certainties. For example, the likelihood of bias in data synthesis is reduced by the use of meta-analysis and the likelihood that a search strategy locates all relevant articles is increased if several electronic databases are used and a rigorous search for grey literature is employed.

5. Conclusions

Given that (i) evidence syntheses are an important means by which policymakers in conservation and environmental science can incorporate scientific consensus into decision-making, (ii) there is an increasing number of policy-relevant evidence syntheses in the field, and (iii) these syntheses exhibit considerable variation in rigour (Fig. 1; Roberts et al., 2006), we believe that CEESAT is a valuable tool in a range of contexts. The standardised guidelines can be used by researchers to objectively identify questions that lack a robust review of all currently available evidence, by editors or peer-reviewers in policy-relevant fields for preliminary screening of manuscripts, and by authors to assist in improving the overall quality of syntheses. Perhaps most importantly, CEESAT assessments conducted by registered assessors could be made available through an open-access database of evidence syntheses, which policymakers would use to identify and compare relevant reviews. We feel that such a resource would be highly valuable in assisting policymakers faced with several reviews on similar topics, and would also promote more effective incorporation of the reliability and rigour of reviews into decision-making.

Acknowledgements

We thank our potential end-users Ally Dingwall (Sainsbury's), Tom Pickerell (Seafish, Seafood Watch), Jon Harman (Seafish), Mike Mitchell, David Parker (Young's Seafood), David Jarrad (Shellfish Association of Great Britain) for their formative feedback. We also thank four anonymous reviewers for helpful and constructive comments. This project was supported by a UK Natural Environmental Research Council Knowledge Exchange Grant NE/J006386/1.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.biocon.2014.04.020.

References

Borenstein, W., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-analysis. John Wiley & Sons, Chichester.
Burda, B.U., Norris, S.L., Holmer, H.K., Ogden, L.A., Smith, M.E.B., 2011. Quality varies across clinical practice guidelines for mammography screening in women aged 40–49 years as assessed by AGREE and AMSTAR instruments. J. Clin. Epidemiol. 64, 968–976.
Branton, M., Richardson, J.S., 2011. Assessing the value of the umbrella-species concept for conservation planning with meta-analysis. Conserv. Biol. 25, 9–20.
CEE, 2013. Collaboration for Environmental Evidence. Guidelines for systematic review in environmental management Version 4.2. Environmental Evidence. <http://www.environmentalevidence.org/Documents/Guidelines4.2.pdf> (accessed 01.07.13).
Cohen, J., 1960. A coefficient of agreement for nominal scales. Edu. Psychol. Meas. 20, 37–46.
CRD, 2009. Centre for Reviews and Dissemination. CRD's guidance for undertaking reviews in health care. CRD, University of York. Available at: <http://www.york.ac.uk/inst/crd/publications.htm> (accessed 21.11.12).
Dent, D.H., Wright, S.J., 2009. The future of tropical species in secondary forests: a quantitative review. Biol. Conserv. 142, 2833–2843.
Englund, G., Samelle, O., Cooper, S.D., 1999. The importance of data-selection criteria: meta-analyses of stream predation experiments. Ecology 80, 1132–1141.
Fazey, I., Salisbury, J.G., Lindenmayer, D.B., Maindonald, J., Douglas, R., 2004. Can methods applied in medicine be used to summarise and disseminate conservation research? Environ. Conserv. 31, 190–198.
Gates, S., 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. J. Anim. Ecol. 71, 547–557.
Halpern, B.S., 2003. The impact of marine reserves: do reserves work and does reserve size matter? Ecol. Appl. 13, 117–137.
Harrison, F., 2011. Getting started with meta-analysis. Methods Ecol. Evol. 2, 1–10.
Huntington, B.E., 2011. Confronting publication bias in marine reserve meta-analyses. Front. Ecol. Environ. 9, 375–376.
Kalies, E., Covington, W., Chambers, W., Rosenstock, S., 2010. How do thinning and burning treatments in southwestern conifer forests in the United States affect wildlife density and population performance? CEE Review 09-005 (SR66). Collaboration for Environmental Evidence: <http://www.environmentalevidence.org/SR66.html>.
Lajeunesse, M.J., Forbes, M.R., 2003. Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecol. Lett. 6, 448–454.
Lajeunesse, M.J., 2010. Achieving synthesis with meta-analysis by combining and comparing all available studies. Ecology 91, 2561–2564.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.
Lawler, J.J., Aukema, J.E., Grant, J.B., Halpern, B.S., Kareiva, P., Nelson, C.R., Ohleth, K., Olden, J.D., Schlaepfer, M.A., Silliman, B., Zaradic, P., 2006. Conservation science: a 20-year report card. Front. Ecol. Environ. 4, 473–480.
Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 6 (7), e1000100.
Maliao, R.J., White, A.T., Maypa, A.P., Turingan, R.G., 2009. Trajectories and magnitude of change in coral reef fish populations in Philippine marine reserves: a meta-analysis. Coral Reefs 28, 809–822.
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 (7), e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097.
Norris, R.H., Webb, J.A., Nichols, S.J., Stewardson, M.J., Harrison, E.T., 2012. Analysing cause and effect in environmental assessments: using weighted evidence from the literature. Freshw. Sci. 31, 5–21.
Peppin, D., Fulé P., Beyers, J., Sieg, C., Hunter, M., 2011. Does seeding after severe forest fire in western USA mitigate impacts on soils and plant communities? CEE review 08–023 (SR60) Collaboration for Environmental Evidence: <http://www.environmentalevidence.org/SR60.html>.
Philibert, A., Loyce, C., Makowski, D., 2012. Assessment of the quality of meta-analysis in agronomy. Agri. Ecosyst. Environ. 148, 72–82.
Popovich, I., Windsor, B., Jordan, V., Showell, M., Shea, B., Farquahar, C.M., 2012. Methodological quality of systematic reviews in subfertility: a comparison of two different approaches. PLoS ONE 7 (12), e50403. http://dx.doi.org/10.1371/journal.pone.0050403.
Pullin, A.S., Knight, T.M., 2001. Effectiveness in conservation practice: pointers from medicine and public health. Conserv. Biol. 15, 50–54.
Pullin, A.S., Knight, T.M., 2003. Support for decision making in conservation practice: an evidence-based approach. J. Nat. Conserv. 11, 83–90.
Pullin, A.S., Knight, T.M., 2009. Doing more good than harm – building an evidence-base for conservation and environmental management. Biol. Conserv. 142, 931–934.
Pullin, A.S., Knight, T.M., 2012. Science informing policy – a health warning for the environment. Environ. Evid. 1, 15.
Roberge, J.-M., Angelstam, P., 2004. Usefulness of the umbrella species concept as a conservation tool. Conserv. Biol. 18, 76–85.
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. Biol. Conserv. 132, 409–423.
Rosenberg, M.S., 2005. The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. Evolution 59, 464–468.
Shea, B.J., Bouter, L.M., Peterson, J., Boers, M., Andersson, B.M., Ortiz, Z., Ramsay, T., Bai, A., Shukla, V.K., Grimshaw, J.M., 2007. External validation of a measurement tool to assess systematic reviews. PLoS ONE 2 (12), e1350. http://dx.doi.org/10.1371/journal.pone.0001350.

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. J. Clin. Epidemiol. 62, 1013–1020.

Stewart, G.B., 2010. Meta-analysis in applied ecology. Biol. Lett. 6, 78–81.

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends Ecol. Evol. 19, 305–308.

The Cochrane Collaboration, 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. In: Higgins, J.P.T., Green, S. (Eds.), Available from <http://www.cochrane-handbook.org>. (updated March 2011).

Trimble, M.J., van Aarde, R.J., 2012. Geographical and taxonomic biases in research on biodiversity in human-modified landscapes. Ecosphere 3 (12), 119, <http://dx.doi.org/10.1890/ES12-00299.1>.

Viera, A.J., Garrett, J.M., 2005. Understanding interobserver agreement: the kappa statistic. Fam. Med. 37, 360–363.

Veríssimo, D., MacMillan, D.C., Smith, R.J., 2011. Toward a systematic approach for identifying conservation flagships. Conserv. Lett. 4, 1–8.

Whittaker, R.J., 2010. Meta-analysis and mega-mistakes: calling time on meta-analysis of the species richness-productivity relationship. Ecology 91, 2522–2533.

Zumsteg, J.M., Cooper, J.S., Noon, M.S., 2012. Systematic review checklist. A standardised technique for assessing and reporting reviews of life cycle assessment data. J. Ind. Ecol. 16, S12–S21.