Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT) criteria and scoring guidelines for reliability of evidence reviews.

For details on appropriate use and limitations of this tool please refer to Woodcock et al (2013) *Biological Conservation* 176, 54-62.

Guidelines for applying each of 13 scoring criteria are given below. Evidence syntheses receive 3 points (GREEN), 1 point (AMBER), or 0 points (RED) for each criterion. Rationale for each criterion is provided , 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'.

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 c*riteria 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.

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.

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.