Guidelines for Systematic Review in

Conservation and Environmental Management.

CENTRE FOR EVIDENCE-BASED CONSERVATION

SCHOOL OF THE ENVIRONMENT & NATURAL RESOURCES

BANGOR UNIVERSITY

BANGOR, GWYNEDD LL57 2UW

UNITED KINGDOM

VERSION 3.1

February 2009

www.cebc.bangor.ac.uk

Preamble

In response to problems of accessing scientific information to support decision-making, many applied disciplines are utilising an evidence-based framework for knowledge transfer involving systematic review and dissemination of evidence on effectiveness of interventions at the practical and policy levels (Stevens & Milne 1997; Khan 2003). The framework is most fully developed in the health services sector, where global review and dissemination units have been established and are linked by networks such as the Cochrane Collaboration (www.cochrane.org). Within these networks, systematic reviews are undertaken following set guidelines, that include peer review, to ensure that they meet required standards before dissemination. The need for such a framework in environmental conservation has been argued elsewhere (Pullin & Knight 2001; Fazey et al. 2004; Pullin et al. 2004; Sutherland et al. 2004). Here we present the latest guidelines for systematic review and dissemination in biodiversity conservation and environmental management.

In the following guidelines we have used established methods from the health services sector (NHS CRD 2001; Khan 2003; Higgins & Green 2005) as our models. By undertaking our own systematic reviews to test these models we have modified the guidelines, through analysis of procedures and outcomes, for their application to conservation and environmental management. Although the basic ethos of systematic review remains unchanged, ecological data are often fundamentally different in nature from data on human health (Fazey et al. 2004; Pullin et al. 2004), and this is reflected in our guidelines. At first glance, many of the guidelines may seem routine and common sense, but the rigour and objectivity applied at key stages, and the underlying philosophy of transparency and independence, sets them apart from the majority of traditional reviews recently published in the field of applied ecology (Roberts et al. 2006). Pullin and Knight (2001), Fazey et al. (2004), Pullin et al. (2004), and Sutherland et al. (2004) argue that, once established, systematic review methodology will significantly improve the identification and provision of evidence to support practice and policy in conservation and environmental management. For this methodology to have an impact on conservation effectiveness, more conservation biologists need to undertake reviews, and we encourage this community to use, and improve, these guidelines and help establish an evidence-based framework for our discipline.

Systematic Review Guidelines

For clarity the guidelines are split into three stages and key phases within each.

1. Planning the review
2. Conducting the review
3. Reporting and dissemination of results

We use examples of our own reviews to highlight key issues.

Contents

Stage 1 - Planning the review

1.1 Question formulation

A systematic review starts with a specific question, clearly defined with subject, intervention and outcome elements (Table 1), that is answerable in scientific terms (Jackson 1980; Cooper 1984; Hedges 1994). The question is critical to the process because it generates the search terms used in the subsequent literature review and determines relevance criteria (NHS CRD 2001). Finding the right question is a compromise (probably more so in ecology than in medicine) between taking a holistic approach, by involving a large number of variables and increasing the number of relevant studies, and a reductionist approach that limits the review's relevance, utility, and value (Stewart et al. 2005). The question should be practice or policy relevant and should therefore be generated by, or at least in collaboration with, relevant decision-makers (or organisations) for whom the question is real. It may also be important for the question to be seen as neutral to stakeholder groups. Ideally meetings should be held with key stakeholders to try and reach consensus on the nature of the question. This may be more critical for ecological review than medical review because, unlike the benefit of improving human health, the benefit of conserving biodiversity is often contested (Fazey et al. 2004).

Table 1. Elements of a reviewable question; normally a permutation of 'does intervention x on subject y produce outcome z'.

Question element	Definition
Subject	unit of study (e.g., ecosystem, habitat, species) that should be defined in terms of the subject(s) to which the intervention will be applied
Intervention	proposed management regime, policy, or action
Outcome	all relevant objectives of the proposed management intervention that can be reliably measured. Particular consideration must be given to the most important outcome and to any outcome that has greater benefits or disadvantages than any alternatives (i.e., the outcome desired)
Comparator	Either a control with no intervention or an alternative intervention

Example of question formulation

Natural England, a UK statutory conservation agency, was concerned about the ecological impacts of burning management carried out by landowners in upland areas of England. Discussion with Natural England personnel enabled this general concern to be "unpacked," allowing definition of subject, intervention, and outcome elements of two specific review questions (Stewart *et al.* 2005): "Does burning of U.K. submontane, dry dwarf-shrub heath maintain vegetation diversity?" and "Does burning degrade blanket bog?" Identification of these two related questions allowed specific hypotheses to be tested whilst retaining broader policy relevance. These also provided examples of habitat-based reviews.

Although discussions with review proposers have proven effective in the formulation of a review question, other stakeholders may disagree. In the above example, a key stakeholder disagreed with the outcome measure (a measure of favourable ecological condition based on the relative abundance of key species) used in the "blanket bog" review. To avoid post-review problems such as this we advocate involvement of multiple stakeholders early in the review process (see 'Developing a Review Protocol' section below)

1.2 Review scoping

Before the commencement of a systematic review, it is essential that review teams undertake a period of review scoping. This phase will guide not only the construction of a comprehensive and appropriate protocol, but will also provide an indication of the likely form of the review and thus facilitate resource planning. A thorough scope mirrors the review process in its stages. Specifically it should entail:

- The construction of a full search strategy
- An assessment of the volume of relevant literature
- A trial critical appraisal of study quality and data extraction

1.2.1 Constructing a search strategy

The development of an effective search strategy will most likely be an iterative process, with sensitivity improving as scoping progresses: all iterations of tested terms should thus be recorded, with the number of 'hits' they return (see Appendix A for an example). This should be accompanied with an assessment of proportional relevance, so that the number of hits returned is not taken alone as an indication of the volume of relevant material. Comparisons of individual terms will allow the identification of superfluous or ineffective keywords, and thus their removal from the search strategy. It is important to remember however, that terms apparently useful in literature databases will not always be appropriate when conducting general web searches and thus parallel strategies may need to be developed. All scoping searches should be saved so that they may be accessed during the search phase of the review (see Stage 2 – Searching for data), removing duplication of effort where possible. However, if the scoping search is conducted well in advance of the actual review search, it would be prudent to conduct the search again in order to ensure all recent literature has been identified.

It is important that the search for literature and data is sufficiently rigorous and broad so that as many studies as possible that are eligible for inclusion are identified. The search strategy is constructed from search terms extracted from the subject, intervention, and outcome elements of the question, combined where appropriate using Boolean operators ('AND', 'OR', 'NOT', etc.), and utilising wildcard truncation symbols to search for variant word endings (see the Example below). This may include considering synonyms, alternative spellings, and non-English language terms within the search strategy. Search protocols must balance sensitivity (getting all information of relevance) and specificity (the proportion of "hits" that are relevant) (NHS CRD 2001). A comprehensive search improves the credibility of the review because more of the relevant studies available are captured by the search and included in the review. In ecology, searches of high sensitivity often come at cost of lower specificity, which means searches are resource-intensive. This is because ecology lacks the MeSH

(Medical Subject Headings)-indexes and integrated databases of medicine and public health, which assign standard keywords/descriptors to articles. A high-sensitivity and low-specificity approach is necessary to capture all or most of the relevant articles available, and reduce bias and increase repeatability in capture (see below). Typically, large numbers of references are therefore rejected. For example, of 317 articles with relevant titles concerning the impact of burning on blanket bog, only 8 (2.5%) had comparators (Stewart *et al.* 2005). Similarly, reviews regarding burning of dry heath and the impact of windfarms on bird abundance resulted in meta-analysis of 1.7% and 12%, respectively, of material with relevant titles.

Where you know of existing meta-analyses or reviews, or subject experts have identified a set of relevant studies, a final step in the development of the search strategy is to use these to test the strategy. A comprehensive strategy with an appropriate balance of specificity and sensitivity will retrieve all of the known relevant studies without returning an unmanageable number of hits.

Example of a search strategy

A review of the effectiveness of control methodologies on introduced populations of the American Mink (*Mustela vison*) in Europe (Tyler et al. 2005) searched 14 electronic databases (Agricola, BIOSIS Previews, CAB Abstracts, Copac, Digital Dissertations, Index to Theses Online, ISI Current Contents, ISI Proceedings, ISI Web of Science, JSTOR, ScienceDirect, Scirus, Scopus, English Nature's Wildlink catalogue); the World Wide Web (first 100 "hits" from www.alltheweb.com, www.google.co.uk, U.K. Department for the Environment, Food and Rural Affairs, Scottish Natural Heritage, Oxford University's Wildlife Conservation Research Unit, The Royal Society for the Protection of Birds, The National Trust, British Wildlife, The Mammal Society, Mammals Trust, and The British Trust for Ornithology); and bibliographies of relevant articles. The search terms used were: *Mustela* AND *vison*, *Mustela* AND *vison* AND trap*, *Mustela* AND *vison* AND control*, *Mustela* AND *vison* AND management, *Mustela* AND *vison* AND pest, Mink AND trap*, Mink AND control*, Mink AND management, Mink AND pest). The specificity of this search was low with considerable overlap between resources (many references were identified multiple times). Specificity could have been increased by using the species name as a search term rather than separating it, i.e. *"Mustela vison" and "M. vison"*. The grey literature search was largely U.K.-based due to resource limitations, although the inclusion of non-U.K. theses was possible. The low specificity of the review (only 1% of retrieved material was judged relevant), however, limits the potential for bias notwithstanding the geographical scope of the grey-literature search. The documented search is fully repeatable and transparent; thus, readers can judge its validity.

1.2.2 Assessing the volume of literature

The volume of literature arising from a scoping search (see above section) may be used as a crude predictor of whether the review will identify a knowledge gap or if it has the potential to provide some form of data synthesis. This has implications in terms of the time and resources required to complete the review and, in some cases, deciding whether it is worthwhile proceeding with a particular review topic.

This stage should also involve the identification of stakeholders and experts, who should be contacted to provide guidance on the range of relevant study methodologies and citations, and the identification of other potential sources of data, whether individuals or organisations.

1.2.3 Trial critical appraisal, data extraction and analysis
Having developed an effective search strategy and become more familiar with the potentially relevant material, the next step in a review scope should be an in-depth examination of a sub-set of the apparently relevant material. This sub-set may comprise of primary studies identified from existing reviews or meta-analyses, a known range suggested by subject experts, and/or those citations in the literature which have formed the rationale for the review. Critical appraisal of a sub-set of relevant articles will enable the identification of the sources of uncertainty within these primary studies.

Having critically appraised a sub-set of relevant articles, it may be possible to perform a pilot data extraction and synthesis. This will inform the development of a suitable data extraction strategy and analysis approach by allowing, for example: the identification of the range of data types and methodological approaches; the determination of appropriate effect size metrics (e.g. meta-analysis or qualitative synthesis?); and the identification of study co-variates. This information, in turn, will feedback into the development of a comprehensive protocol and facilitate resource planning.

Whilst the process of scoping may seem like a time-consuming one, the benefits of a properly conducted scope are considerable and this early investment will doubtless be paid back several-fold by improved focus and efficiency throughout the later stages of the review.

1.3 Developing a review protocol

The review protocol acts as a document that all stakeholders agree upon, after which the review itself can be conducted. Consider convening a stakeholder meeting at this stage to seek consensus on the review question and the conduct of the review. It is preferable to be aware of stakeholder conflicts at this stage than to discover them later.

The protocol makes clear what the review relates to, and is useful for getting the engagement of experts who may have data to contribute. Anyone reading the protocol should clearly understand what the question is and what data are required. To satisfy the philosophy of transparency in undertaking systematic review, the draft review protocol should be made openly available for comment (e.g. on the Collaboration for Environmental Evidence (CEE) website for a one month period), enabling others who have not been contacted during the development stage to provide comments on the direction of the review. The protocols are made available on the CEE website to show which reviews are in progress, enabling others to see if a review is being conducted that they may be interested in, or to prevent starting a review on a topic that is already underway (see www.environmentalevidence.org for examples). A review protocol can also be used to determine the amount of resources required to conduct a review, whether people, time or money, and to allocate activities to different members of a review team.

A review protocol is developed as a document that guides the review. As in any scientific endeavour, methodology should be established and made available for scrutiny and comment at an early stage. Because reviews are retrospective by nature, the protocol is essential to make the review process as rigorous, transparent, and well defined as possible (Light 1984). Beside a formal presentation of the question and its background (the “real world” context), a review protocol sets out the strategy for obtaining primary data and defines relevance criteria for data inclusion or exclusion (NHS CRD 2001). The subject, intervention and outcome elements defined in the question-setting stage provide *a priori* inclusion criteria important for the objectivity and transparency of the review. If the relevant population, intervention, or outcome measures are present, then data are included. The protocol should also establish the methods to be used for critical appraisal, data extraction and synthesis, and state any conflicts of interest in the review plus sources of funding. For guidance on developing a review protocol go to www.environmentalevidence.org/Authors.htm

By planning the review in advance, the protocol helps minimise bias within the review. It may become necessary during the course of a review to make changes to the protocol. These changes should be clearly documented within the final review so that transparency and repeatability is maintained.

Stage 2 - Conducting the review

2.1 Searching for data

A wide range of sources should be accessed to capture information. The primary method for information retrieval is the systematic literature search, but this should be supplemented by the checking of bibliographies, the provision of supplementary data from authors and through contact with subject experts. Different questions may require the use of different resources, and searches may produce different types of results depending on the information available. Databases and catalogues vary in the manner in which they can be searched. Searches may often have to be modified between resources as a consequence. Database help files can be useful to ascertain the search capabilities, such as the symbols for wild card terms and the use of parenthesis and Boolean terms. Many of the well-known databases allow complex search strings. However, others only allow searching with single keywords but should still be included in the search. Obviously resource availability will constrain the numbers of literature sources used and search term permutations applied, which will also be subject to diminishing return. Managing the citations within a bibliographic software package can be useful to assess the amount of duplication in articles captured as the search proceeds. It is important to record the methods used in all parts of the search so that others can judge the probability that important research has been missed and so that transparency and repeatability is maintained (see “Developing a review protocol”)

The literature search should be comprised of three distinct phases:

1. searching online databases and catalogues
2. searching organisations and professional networks
3. searching the web.

2.1.1 Searching online databases and catalogues

There are a number of general scientific electronic databases that may be useful for identifying relevant articles and data sets, such as Web of Science and Scopus. Access to most of these depend on Library subscriptions, and so varies between institutions and organisations. If you have one, contacting the subject librarian to identify and discuss the resources available within your organisation is recommended at an early stage of the protocol development. As well as the general scientific databases, there are also some subject-specific databases that may contain relevant information, and it may be necessary to search local databases for questions with a regional focus.

Different databases and catalogues sample different subsets of the literature, and so multiple sources should be accessed to ensure the search is unbiased. To ensure the search is comprehensive yet practical, it can be useful to consider the slant of each database and catalogue to ensure that at least one resource is searched to cover different subsets of the literature, for example, theses and dissertations, peer-reviewed and non-peer-reviewed published articles and so-called 'grey literature' that has not been formally published.

Organisations often have access to different resources, and so the list of resources searched for each review will vary, but checking bibliographies and contact with authors should help to ensure all references are retrieved. To minimize the problem of publication bias (e.g., Leimu & Koricheva 2005), both published and unpublished data must be included, a standard rarely satisfied in traditional reviews. The next two stages of the literature search help to address this issue.

2.1.2 Searching organisations and professional networks

Many organisations and professional networks make documents freely available through their web pages, and many more contain lists of projects, datasets and references. Often, reports referred to on a website will be provided if an organisation is contacted. Sometimes, a visit may be necessary when a large number of documents are required. Searching these organisations and networks targets the grey literature which would not come up in a conventional database search. The list of organisations to be searched is dependent upon both the subject of the systematic review and any regional focus.

If feasible, hand searching of specific sources and visits to institutions (e.g. libraries and museums) may be advantageous to extract all relevant material.

2.1.3 Web searching

The Internet can be a useful tool for identifying unpublished and ongoing studies, as well as locating subject experts and relevant organisations. Careful consideration must be given to the design of the search in order to ensure that it is as focused and specific as possible (Eysenbach 2001); where this is not done, searching the web can be a time-consuming task, with relatively little useful data being returned. Thus, scoping (see above) should form a key component of any web searching strategy; piloting of potential search terms is essential, as any ambiguity is likely to return a larger proportion of spurious results. An awareness of differences in search engine functionality is also important, as these may impose inconsistencies in approach but it is reasonable to tailor the search to the search engine to maximise its usefulness.

The indexable web is now some several billion pages in size and, whilst a wide range of engines exist to enable users to search these pages, none of these individually index more than approximately 30% of the total web (Lawrence & Giles 1998). Overlap studies (e.g. Dogpile 2007) suggest that there is relatively little cross-over between the major search engines, with the proportion of results unique to each engine as high as 88%. Therefore, to ensure maximum retrieval of the available relevant information, it is essential that multiple engines are searched. The use of meta-engines, such as Excite, Vivismo, Kartoo and Dogpile, which simultaneously search a number of individual engines, may also offer a part-solution to the problem of patchy coverage. In general meta-engines should be treated with caution as many of these search only the free, poorer quality engines and, in cases where the most useful engines are included, limits on the number of hits returned from each engine often mean that such searches are considerably less useful than individual searches of the single engines (University of California 2008).

In addition to discrepancies in the extent of individual web coverage, there are disparities in the ways in which search engines rank their results. Page position within the results is not necessarily correlated to the relevance or quality of the documents retrieved. Although a closely-guarded secret, the ranking algorithms employed by major search engines are primarily based on one or more of a set of general principles. Most use the frequency and location of keywords as a fundamental guide of relevance, with those pages containing the specified search terms most frequently and higher up the document, appearing at the top of the results listing (Hock 1999). Others determine relevance from a 'popularity' scoring system, whereby pages are ranked according to the number of sites that link to them, with high rankings associated with high 'link popularity'(Introna 1999). The majority of search engine providers effectively sell search positions in one form or another: most differentiate these 'sponsored' results from 'standard' ones, but it is not uncommon for the former to be embedded within the main results page and be otherwise indistinguishable from the latter. Issues with engine ranking systems will become clear during the scoping phase, and should be used to guide decisions as to engine inclusion into the final review search strategy.

Boolean logic (the use of "AND", "OR", "NOT" indicators) is supported in varying degrees by the major search engines, as is truncation using wildcards. These capabilities can be checked in the engine's accompanying 'help' files when selecting engines for inclusion. Many engines lack a nesting feature (use of parentheses) that would enable the use of more complex Boolean queries (Hock 1999). Where the nature of the study necessitates multi-element search strings, it may be possible to reconstruct these searches using the advanced search features offered: the majority of the major search engines provide a "find all the words" and "find any of the words" feature which is particularly helpful.

When searching the internet for grey literature, it might be more efficient to run searches with a restriction on the file type to be returned, on the premise that these may be more likely to contain useful data than standard web pages. For example, by limiting the search to Excel spreadsheets, raw data that would otherwise have ranked low in an unrestricted search may be captured. Most search engines provide the option of file restriction to a range of formats (.pdf, .doc, .xls, .rtf, etc.) and this is usually accessed via the engine's "advanced search" page. A small number of engines (e.g. Scirus) allow the selection of multiple file formats per search: most do not however, and where this is desired, visual sorting of the search results may be the only solution. Searches such as these should be recorded as part of the search strategy.

In addition to the more general search engines, the incorporation of specialised subject gateways searches into web searching strategies may be helpful. Databases such as Intute.ac.uk, ScienceResearch.com and AcademicInfo.net, contain links to hand-selected sites of relevance for a given topic or subject area and are particularly useful when searching for subject experts or pertinent organisations, helping to focus the searching process and ensure relevance.

Specific guidance on how much searching is acceptable is difficult to give. In the medical literature, papers sometimes cite a "first 50 hits" approach (e.g. Smart & Burling 2001), whereby the first 50 results for each search are viewed in full. However, this appears to be an arbitrary number, and is more likely based upon the resources available to the review team than a reflection of the extent of searching required to effectively capture the most-relevant grey literature available. Given that the actual number of hits retrieved is review-specific, related both to the search terms used and the quantity of information available, in some instances there may be a case for modifying the recommended search limits (e.g. if there are particularly large or small numbers of relevant hits). Thus, in order to provide a consistent and practical way to limit web searching, we would recommend, at a minimum, the full viewing of each of the first 50 hits. The proportion of relevant material retrieved in this subset will then provide an indication as to the potential utility of examining further hits. Review teams must also decide the extent to which links from the original 'hits' to potentially relevant material will be followed, and must make sure the chased links are recorded in each instance (if a pre-determined limit is not set). It is important, both for citation purposes (should an online document be selected for inclusion in the review) and to ensure transparency and repeatability, that the dates of the web searching phase are clearly documented: the use of a simple recording form will facilitate this.

2.2 Selection of relevant data

Once searching is complete, relevant articles must be efficiently selected without wasting resources examining irrelevant articles in too much detail. Selecting only relevant articles from a potentially large body of initial literature requires the reviewer to use inclusion and exclusion criteria stated *a priori* in the protocol. These criteria relate directly to the elements of the question (subject, outcome, intervention and comparator; Table 1).

These criteria can be applied at different levels of reading (title, title and abstract and full text) to impose a number of filters of increasing rigor. If a long list of articles or data sources is acquired (1000s rather than 100s) and the list of relevant sources is likely to be much shorter, it may be efficient to exclude some material on title only, especially if obviously spurious hits arise from ambiguity in the search words. The second filter should examine the title and abstract to determine relevance. However, the approach should be conservative so as to retain articles if there is reasonable doubt as to whether all the inclusion criteria are met. For instance, on reading title and abstract, it is often difficult to assess whether a study has the relevant comparator. If there is no abstract then the article should be retained.

It is good practice at, the beginning of the abstract assessment stage, for two reviewers to go through the same process on a random sub-sample of articles from the original list (the recommended sample is a minimum of 20% up to a maximum of 1000 references). To check for consistency in the interpretation of the selection criteria, reviewer relevance decisions can be compared by performing a kappa analysis, which adjusts the proportion of records for which there was agreement by the amount of agreement expected by chance alone (Cohen 1960; Edwards 2002). A kappa rating of 'substantial' (0.5 or above) is recommended to pass the assessment. If comparability is not achieved, then the criteria should be further developed by redefining the scope and interpretation of the question elements. Ideally kappa should be repeated on a new sample of articles, if resources allow, to check the accuracy of the redefined criteria.

Remaining articles, which have not been excluded after reading their titles and abstract, should be viewed in full to determine whether they contain relevant and usable data. Independent checking of a sub-sample by kappa analysis can be repeated at this stage. Obtaining the full text of all articles can be very time consuming and a realistic deadline may have to be imposed and a record kept of those articles not obtained. Short lists of relevant articles and datasets should be made available for scrutiny by stakeholders and subject experts. All should be invited, within a set deadline, to identify relevant data sources they believe are missing from the list. Reviewers should be aware that investigators often cite selectively studies with positive results (Gotzsche 1987; Ravnskov 1992); thus, checking bibliographies and direct contacts must be used only to augment the search.

2.3 Assessing quality of methodology

The quality of the studies included into a systematic review is of critical importance to the resulting quality of the review; if the data are of poor quality then the conclusions cannot be considered robust. In an ideal world, each data set included in a systematic review should be of high methodological quality, thus ensuring that the potential for error and bias is minimized and that any differences in the outcome measure between experimental groups can be attributed to the intervention. To determine the level of confidence that may be placed in selected data sets, each one must be critically appraised to assess the extent to which its research methodology is likely to prevent systematic errors or bias (Moher 1995).

In the health services a hierarchy of research methodology is recognized that scores the value of the data in terms of the scientific rigor (Stevens & Milne 1997). The hierarchy of methodological design can be viewed as generic and has been transferred from medicine to ecology (Pullin & Knight 2003). Where a number of well-designed, high-quality studies are available, others with inferior methodology may be demoted from subsequent quantitative analysis to qualitative tabulation, or rejected from the systematic review entirely. However, there are dangers in the rigid application of this hierarchy to ecology as the importance of various methodological dimensions within studies will vary, depending on the study system to which an intervention is being applied. For example, a rigorous methodology, such as a randomized controlled trial (RCT), , applied over inadequately short time and small spatial scales could be viewed as superior to a time series experiment providing data over longer time and larger spatial scales that were more appropriate to the question. The former has high internal validity but low external validity or generalisability in comparison to the latter. This problem carries with it the threat of misinterpretation of evidence. Potential pitfalls of this kind need to be considered at this stage and explored in covariate analyses (e.g., experimental duration or study area: see Downing et al. 1999 and Côté et al. 2001 respectively) or by judicious use of sensitivity analysis (see below).

Four sources of systematic bias that may threaten the internal validity of a study are routinely considered in healthcare (Feinstein 1985; Moher 1995; Moher 1996; Khan 2003). Selection bias results from the way that comparison (e.g., treatment and control) groups are assembled (Kunz 1998) and is a primary reason for randomization in studies. This is common in conservation ecology because interventions or treatments are applied to valuable sites and analogous controls often do not exist e.g. marine protected areas.

Performance bias refers to systematic differences in the care provided to subjects in the comparison groups and is dealt with by the experimenter being unaware of which are treatments and which controls (blinding) (Shultz 1995). We postulate that the ecological equivalents of performance bias arise from biased baseline comparisons i.e. unequal balancing of heterogeneity in treatment and control arms and failure to consider the impact of covariables that may confound the effectiveness of the intervention. However, it is not possible to account for the influence of potentially confounding variables that are not known or were not measured. Even for those that have been identified, difficulties can arise in extracting standardised information for analysis.

Measurement or detection bias refers to systematic differences incurred when knowledge of the intervention influences the assessment of the results in the comparison groups and is also addressed by blinding (Shultz 1995). Blinding is generally not possible in ecology and the extent of detection bias will therefore vary, depending on the rigour and objectivity of sampling methodology (e.g., percent cover assessed by eye is subject to greater potential detection bias than frequency).
The fourth, attrition bias (systematic differences between the comparison groups in the loss of samples) is common in medical analyses e.g. patients who die are excluded from an outcome group. This can be addressed by analysing all the data, but access to raw data may be a pre-requisite to quantify the impact of attrition bias.

Assessing the quality of methodology is a critical part of the systematic review process. It requires a number of subjective decisions about the relative importance of different sources of bias and data quality elements specific to ecology, particularly the appropriateness of variable temporal and spatial scales. It is therefore vital that the assessment process be standardized and as transparent and repeatable as possible. At least 25 scales and 9 checklists have been used to assess the validity of randomized controlled trials in medicine (Moher 1995; Moher 1996). Juni et al. (1999) evaluated 17 health care trials from a meta-analysis, using these 25 different methodological quality scales. For 12 of the scales, the outcomes of the trials were comparable. However, for 6 scales, high quality trials showed little or no benefit of treatment compared to low quality trials, whilst for the remaining 7 scales the opposite trend was observed. Quality scales can therefore give very different results depending on the data quality items considered and the relevant importance assigned to each one. Similar criteria have also been used to critically appraise the validity of observational studies (Horwitz 1979; Feinstein 1982; Levine 1994; Bero 1999). These checklists do not consider specific ecological criteria. We therefore suggest that review-specific *a priori* assessment criteria for assessing the quality of methodology is included in the protocol and two or more assessors should be used to assess study quality in ecological reviewing. The subjective decisions may be a focus of criticism; thus, we advocate consultation with the scientific community and relevant stakeholders before moving on to data extraction. Pragmatic grouping of studies into high, medium and low quality based on simple but discriminatory checklists of “desirable” study features may be necessary if sample sizes are small and do not allow investigation of all the study features individually.

Finally, at this stage, it may be necessary to reject articles that are seemingly relevant but do not present data in extractable format (e.g., if they do not report standard deviations for control and treatment group(s) or the information required to calculate the statistic). If possible, authors of such articles should be contacted and asked whether they can provide data in suitable format. Contacting authors for data is not normal practice in ecology and can be met with surprise and indignation but it is important to develop the culture and expectation of data accessibility, particularly when the research was publicly funded.

Examples of study quality / methodology assessment

Stewart et al. (2005) used the hierarchy of methodology (Pullin & Knight 2003) to separate randomized controlled trials and site comparisons addressing the question: “Does burning degrade blanket bog?” This reflected a major data-quality divide; therefore, further data-quality assessment was inappropriate given the very small number of studies. This approach enabled a simple, but discriminatory, vote count of studies with results showing positive, neutral, or negative effects.

When reviewing the impact of windfarms on bird populations, the standard hierarchy of evidence was considered inadequate by itself due to variation in other critical data-quality elements. This particularly related to the widespread occurrence of confounding factors resulting from variation between treatment and control at baseline or from changes concurrent with windfarm operation (ecological performance bias). The rigour of observations was also variable as measured in terms of replication and objectivity (ecological detection bias). To test for the impact of these factors, data-quality scores, summing the different aspects of data quality outlined above, were added as a meta-regression covariable. Data-quality score was not significant, suggesting that bifurcation of the data into high- and low-quality evidence was not necessary, possibly because the low-quality studies (low replication, imprecise estimates of abundance, high intratreatment variation coupled with confounded baselines) had a high variance and therefore a low weighting in meta-analysis by inverse variance. Sensitivity analyses were used to explore the impact of including low-quality unreplicated data, but the impact of individual data quality elements other than time was not examined because a large number of environmental and windfarm correlates were of interest and the potential for Type II errors would have been increased. Although this pragmatic approach is easy to apply, there is no measure of a studies' “true” validity (Emerson 1990; Schulz 1995; Jüni 1999). Caution should be exercised in interpreting study validity, especially if different quality elements are combined in a single data-quality sum.

A review of the effectiveness of *Rhododendron* control methods considered study hierarchy and potential for bias providing a subjective summary of data quality (Table 2). In this instance the number of environmental variables with sufficient data for analysis was low and sample sizes were sufficient to examine the impact of some individual study quality variables, such as length of experiment and whether results were generated in the field or a glass house. There were statistically significant differences in effectiveness of control, with greenhouse trials showing greater control than field-based experimentation or monitoring, raising questions about the ecological relevance of greenhouse work and the likely modifying variables. This approach has the merit of objectivity, although there is choice regarding which variables are included in the analysis and caution must be exercised to avoid Type II errors, data mining and overinterpreting results, especially when sample sizes are small.

Table 2. Data quality assessment of an article included in a systematic review of the effectiveness of methods for the control of *Rhododendron ponticum* (Tyler et al. 2004).

Methods	site comparison based on sites treated with different interventions, no control, comparison methods only
Population	no stand-age detail, site located on lowland heath
Intervention and cointerventions	drilled holes filled with herbicide, compared with stumps painted with herbicide
Outcomes	painted stumps 30-40% kill drilled holes 95% kill
Study design	site comparison
Baseline comparison	no information regarding the sites prior to treatment, thus not possible to validate baseline
Intratreatment variation	no information describing intratreatment variation
Measurement of intervention and cointerventions	no information regarding the sites provided, thus not possible to comment on other management within the area
Replication & parameter of abundance	no replication or measure of abundance other than percent kill
Notes	study appears to comment on use of techniques rather than providing the reader with scientific evidence, resulting in a high potential for bias and subsequently low data quality

2.4 Data extraction

Data extracted from articles should be recorded on carefully designed spreadsheets and undertaken with synthesis in mind. Narrative synthesis requires the construction of tables that provide details of the study or population characteristics, data quality, and relevant outcomes, all of which are defined *a priori*. A summary of methodology *in lieu* of study quality assessment may be sufficient where reviews simply summarise available evidence. However, objective qualitative synthesis requires more formal study quality assessment. In such instances data regarding methodology should be extracted to inform critical appraisal in a standardized, transparent and repeatable manner.

Quantitative analysis follows the same model but care must be taken to extract information pertinent to subsequent analysis (e.g., should binary or continuous outcomes be extracted?). In contrast to medicine, consideration of the appropriate spatial scale(s) and level of replication are necessary prior to extracting the variance measures required to weight meta-analyses. Great care must be taken to standardize and document the process of data extraction, the details of which should be recorded in tables of included studies to increase the transparency of the process. To some extent data extraction can be guided by *a priori* rules, but the complexity of the operation means a degree of flexibility must be maintained. Sensitivity analyses can be used to investigate the impact of extracting data in different ways when there is doubt about the optimum extraction method.

In many cases, the information required is not presented and cannot be obtained from authors. In some cases data can be substituted without problems. For example, it is relatively straightforward to substitute standard deviation for standard errors, confidence intervals, *t*-values, or a one-way *F*-ratio based on two groups (Lipsey and Wilson 2001, Deeks *et al*. 2005).

Where missing data cannot be substituted, it can be imputed by various methods. Imputation is a generic term for filling in missing data with plausible values. These are commonly derived from average or standardized values (Deeks *et al*. 2005), but also from bootstrapped confidence limits (Gurevitch & Hedges 2001) or predicted values from regression models (Schafer 1997). Alternatively, data points can be deleted from some analyses, particularly where covariates of interest are missing. Such pragmatic imputation or case deletion should be accompanied by sensitivity analyses to assess its impact.

The impacts of imputation or case deletion can be serious when they comprise a high proportion of studies in an analysis. Case deletion can result in the discarding of large quantities of information and can introduce bias where incomplete data differs systematically from complete (Schafer 1997). Likewise, imputing average values or predicted values from regressions distorts covariance structure resulting in misleading p values, standard errors and other measures of uncertainty (Schafer 1997). Where more than 10% of a data set is missing serious consideration should be given to these problems. More complex imputation techniques are available (see Schafer 1997) and should be employed in consultation with statisticians. If this is not possible, the results should be interpreted with great caution and only presented alongside the sensitivity analysis.

It is difficult to perform formal kappa analysis on the repeatability of data extraction, but some attempt to verify repeatability should be made. A second reviewer should check a random subset (recommended sample of minimum 25%) of the included studies to ensure that the *a priori* rules have been applied or the rationale of deviations explained. This also acts as a check on data hygiene and human error (e.g. misinterpretation of a standard error as a standard deviation). Where data extraction has limited repeatability it is desirable to maintain a record of exactly how the extraction was undertaken on a study by study basis. This maintains transparency and allows authors and other interested parties to examine the decisions made during the extraction process. Particular attention should be paid to the data used to generate effect sizes. Such data extraction forms should be included in an appendix.

Example of data extraction

Reviewing the impact of burning on the ecological condition of blanket bog required extraction of data showing changes in floristic composition and structure. Two reviewers extracted data after reaching a consensus regarding which subsets were relevant within the full data set of each article. *A priori* rules increased the repeatability of data-set formation. For example, sites within an experiment were pooled to prevent pseudoreplication, avoiding post hoc justifications for deriving more than one data-set from an experiment and combining unreplicated, pseudoreplicated and replicated data. Pooled treatment and control sites were included once to maintain independence and avoid bias, with the exception of data on rotational burning, which was scarce and therefore admitted to the review provided there was a comparator irrespective of further potential for bias. Where there was a choice of times since burning, priority was given to the longest time range to maintain independence and maximize predictive power. Similarly, grazed sites received priority over ungrazed sites when the maintenance of independence demanded a choice because grazing and burning are carried out concurrently over most of the British uplands (Stewart *et al.* 2005). If sample sizes had been larger and a quantitative generic outcome measure identified, the impact of these decisions could have been explored with sensitivity analyses. Given the nature of the data, qualitative discussion of the issues was more appropriate.

An example of a data extraction form from a review examining the impact of instream devices on salmonids is shown in Figure 1 overleaf. Note the raw data from which effect sizes were calculated, reference to data sources and information about decisions regarding which data to extract to maintain independence.

Reference	**Binns & Remmick (1994)**					
Location	Huff Creek, Idaho, USA					
Subject	*Oncorhynchus clarki utah* (Bonneville cutthroat trout)					
Intervention	instream habitat structures (36 wooden dams, 9 rock plunges, wooden double deflector, rock deflector, 14 small rock grade controls) rock riprap, fencing of banks					
Methodology	Before and after monitoring					
Sources of bias	Confounding impacts concurrent with the habitat improvement are probably the most important sources of bias. Post improvement droughts occurred resulting in a likely under-estimate of effectiveness.					
Outcomes	post intervention			pre intervention		
	n	m	sd	n	m	sd
Habitat quality index (HQI)	6	38	2	6	30	2
Trout numbers	6	170	59	6	35	18
Reasons for heterogeneity	Monitoring time 11 years. Discharge is extremely variable with a mean of 6ft3/s, stream gradient (1%), proportion of cobbles in					

	substrate (common in half of river, estimated at 25%),degree of existing modification (heavy grazing but river unmodified- low), distance from source (6km), water quality (no information), size of stream (small stream >5m), canopy cover (low >5%).
Pop/pref Extraction	habitat quality pre and post treatment, from text and figure 6. trout numbers from text and table 2. n is the number of sites. Maximum time range was used for post treatment assessment (11 years). Some data is presented for individual sites which allow some separation of features. This was not extracted i) to maintain independence, ii) because no pre treatment assessments are available at a site level
Notes	HQI was evaluated for cut throat trout. Population sizes were estimated using electrofishing (Armour et al. 1983) with degree of population fluctuation assessed as in Platts & Nelson (1988). Much other data regarding both physical habitat and trout was presented but not extracted.
References	Armour, C.L., Burnham, K.P., and Platts, W.S. (1983) Field methods and statistical analysis for monitoring small salmonid streams. U.S. Fish and Wildlife Service FWS/OBS 83/33. Platts, W.S. and Nelson, R.L. (1988) Fluctuations in Trout populations and their implications for land-use evaluation. North American Journal of Fisheries Management 8. 333-345.

Figure 1. Example of a data extraction form (Stewart et al. 2006)

2.5 Data synthesis

This stage includes both qualitative synthesis and quantitative analysis with statistical methods as appropriate. Qualitative synthesis allows informal evaluation of the effect of the intervention and the manner in which it may be influenced by measured study characteristics and data quality. Data from the data-extraction spreadsheet is tabulated to form a summary of the number of data sets providing a yes, no, or neutral answer to each question (vote counting). Where the internal validity of studies varies greatly, reviewers may wish to give greater weight to some studies than others. In these instances it is vital that the studies have been subject to standardised *a priori* critical appraisal with the value judgments regarding internal validity clearly stated. Ideally these will have been subject to stakeholder scrutiny prior to application. More advanced methods for qualitative research synthesis exist and the involvement of specialists should be sought if robust qualitative syntheses are desired (Petticrew & Roberts 2006).

Quantitative analysis can be undertaken to generate overall point estimates of the effect size and to analyse reasons for heterogeneity in the effect of the intervention where appropriate data exist. Meta-analysis is now commonly used in ecology (e.g., Arnqvist & Wooster 1995; Osenberg et al. 1999; Gurevitch & Hedges 2001; Gates 2002); consequently we have not treated it in detail here. Meta-analysis provides summary effect sizes with each data set weighted according to some measure of its importance, with more weight given to large studies with precise effect estimates and less to small studies with imprecise effect estimates. Generally each study is weighted in inverse proportion to the variance of its effect. Pooling of individual effects can be undertaken with fixed-effects or random-effects statistical models. Fixed-effects models estimate the average effect and assume there is a single true underlying effect, whereas random-effects models assume there is a distribution of effects that depend on study characteristics. Random effects models include inter-study variability (assuming a normal distribution); thus, when there is heterogeneity, a random-effects model has wider confidence intervals on its summary effect than a fixed-effect model (NHS CRD 2001; Khan 2003). Random-effects models or mixed models (containing both random and fixed effects) are often most appropriate for the analysis of ecological data because the numerous complex interactions common in ecology are likely to result in heterogeneity between studies or sites. Exploration of heterogeneity is often more important than the overall pooling from a management perspective, as there is rarely a one size fits all solution to environmental problems.

Relationships between differences in characteristics of individual studies and heterogeneity in results can be investigated as part of the meta-analysis, thus aiding the interpretation of ecological relevance of the findings. Exploration of these differences is facilitated by construction of tables that group studies with similar characteristics and outcomes together. Data sets can be stratified into subgroups based on populations, interventions, outcomes, and methodology. Important factors that could produce variation in effect size should be defined *a priori* (see Section 1.3 above) and their relative importance considered prior to data extraction to make the most efficient use of data. Differences in subgroups of studies can then be explored.

If sufficient data exist, meta-analysis can be undertaken on subgroups and the significance of differences assessed (see Box 1.). Such analyses must be interpreted with caution because statistical power may be limited (Type I errors possible) and multiple analyses of numerous subgroups could result in spurious significance (Type II errors possible). Alternatively, a meta-regression approach can be adopted whereby linear regression models are fitted for each covariate, with studies weighted according to the precision of the estimate of treatment effect in a random-effects model (Sharp 1998).

Despite the attempt to achieve objectivity in reviewing scientific data, considerable subjective judgment is required when undertaking meta-analyses. These judgements include decisions about choice of effect measure, how data are combined to form datasets, which data sets are relevant and which are methodologically sound enough to be included, methods of meta-analysis, and the issue of whether and how to investigate sources of heterogeneity (Thompson 1994). Reviewers should state explicitly and distinguish between the *a priori* and *post hoc* rationales behind these decisions to minimize bias and increase transparency.

Quantitative research synthesis is still in its infancy. The biases associated with specific techniques are not generally based on empirical evidence, particularly as applied to ecological research. There is considerable potential to improve the statistical models and to provide robust guidance about which models are most relevant under which circumstances. Pending these developments, we advise reviewers to search for the broad patterns contained in accumulated ecological knowledge using *a priori* decisions and a pragmatic design wisdom to build repeatable knowledge structures with as much structural integrity as possible. Those who are interested in more complex meta-analytical problems may wish to implement meta-analysis using a Bayesian or empirical

Bayesian approach, which allows combination of disparate types of evidence and expresses uncertainty without overfitting data (Gelman et al. 1995).

Example of data synthesis

A review of the impact of wind turbines on bird abundance utilised standardized mean difference meta-analysis with weighting by inverse variance to combine data from 19 globally distributed windfarms. Sensitivity analyses were used to explore the effect of including data from unreplicated studies and to assess bias arising from data extraction of pseudoreplicated or aggregated data. Pooled effect sizes remained negative and statistically significant regardless of how the effect sizes were generated, indicating that the patterns in the data were robust. *A priori* and *post hoc* reasons for heterogeneity were explored with meta-regression. Of the *a priori* variables only bird taxon appeared to modify the result, with relationships between turbine number and power being too weak to have biological significance. *Post hoc* analysis revealed that the impact of windfarms became more pronounced over time, a finding not reported by any of the original research or previously assessed in the literature. This has important implications because declines in local bird abundance are more likely to have deleterious population-level impacts if they worsen over time. It also suggests that current windfarm monitoring programs are of inadequate duration to detect deleterious effects.

Box 1. Interpretation of Forrest plots-Example using STATA

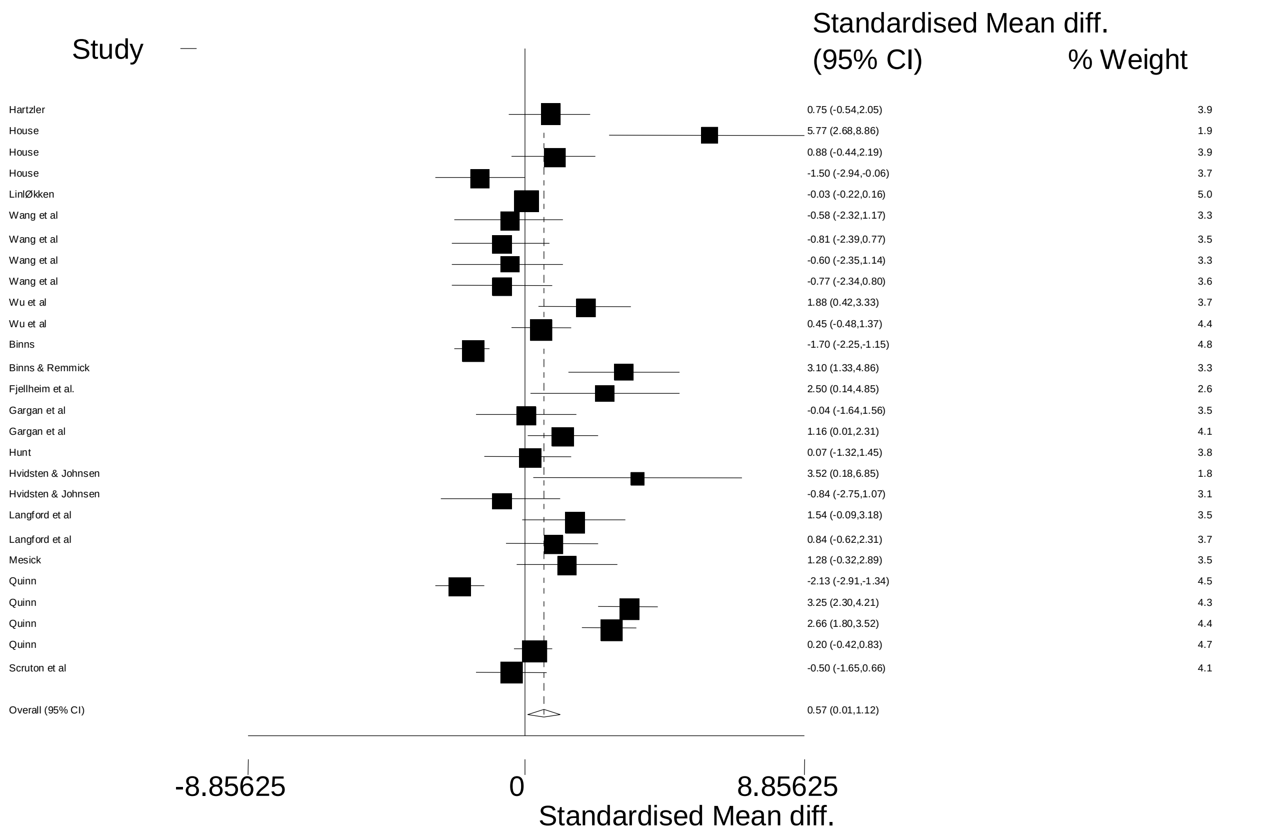

Figure 2. An example of a Forrest plot generated using STATA and typically included as an outcome in reviews that incorporate a meta-analysis.

The individual data points included in the meta-analysis are listed down the left side of the diagram. In this example multiple independent points have been extracted from the same references. Individual studies are typically identified by author name and year, with multiple points numbered. Full details of each study can be found in the references at the end of the systematic review and the tables of included studies and data extraction appendices should make it clear how multiple points were derived from individual studies.

Each data point extracted from a study is represented by a square. The size of the square represents the sample size of that data point whilst the error bar typically represents the 95% confidence interval. The position of the square on the x axis denotes the effect size (in this example Cohens D). This example also lists the effect size and confidence interval for each study to the right of the diagram, along with the weight which that study contributes to the overall synthesis (in this example weighting is by inverse variance).

Underneath the studies, there is a pooled estimate of effect represented by an open diamond. This is a graphical representation of the combined outcome for all of the included data points. The width of this diamond represents the confidence interval.

The “line of no effect” where the effect size is zero is represented by a solid vertical line, and anything that crosses this line is not statistically significant (including those studies where only the confidence interval crosses the line). Anything that falls to the left of the line of no effect has less of the outcome; whereas anything that falls to the right has more of the outcome- whether this is a positive or negative result depends on what the outcome of the meta-analysis is. Therefore a beneficial result for a negative outcome (such as habitat loss) has a significant effect size to the left of the vertical line and a beneficial result for a positive outcome (such as increase in suitable habitat) has a significant effect size to the right of the vertical line. Overall interpretation of the Forrest plot relies on consideration of the position and significance of individual points as well as the pooled estimate, because the pooled estimate can be misleading when heterogeneity is high (see above).

2.6 The interpretation of meta-analysis and systematic review evidence

Systematic reviews synthesise and present evidence but the strength of this evidence and the applicability of the results require careful consideration and interpretation. The discussion and conclusions may consider the implications of the evidence in relationship to practical decisions, but the decision-making context may vary, leading to different decisions based on the same evidence. Authors should, where appropriate, explicitly acknowledge the variation in possible interpretation and simply present the evidence rather than offer advice. Recommendations that depend on assumptions about resources and values should be avoided (Khan 2003, Deeks et al. 2005).

Deeks et al (2005) offer the following advice of relevance here. Authors and end-users should be wary of the pitfalls surrounding inconclusive evidence and should beware of unwittingly introducing bias in their desire to draw conclusions rather than pointing out the limits of current knowledge. Where reviews are inconclusive because there is insufficient evidence, it is important not to confuse 'no evidence of an effect' with 'evidence of no effect'. The former results in no change to existing guidelines, but has an important bearing on future research, whereas the latter could have considerable ramifications for current practice or policy.

Review authors, and to a lesser extent end-users, may be tempted to reach conclusions that go beyond the evidence that is reviewed or to present only some of the results. Authors must be careful to be balanced when reporting on and interpreting results. For example if a ‘positive’ but statistically non-significant trend is described as ‘promising’, then a ‘negative’ effect of the same magnitude should be described as a ‘warning sign’. Other examples of unbalanced reporting include one-sided reporting of sensitivity analyses or explaining non-significant positive results (e.g. the included studies were too small to detect a reduction in mortality for a statistically non-significant increase in mortality) but not negative ones. If the confidence interval for the estimate of difference in the effects of interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well and both should be given equal consideration in discussion of results. One-sided attempts to explain results with reference to indirect evidence external to the review should be avoided. Medical guidance suggests that considering results in a blinded manner can avoid these pitfalls (Deeks *et al.* 2005). Authors should consider how the results would be presented and framed in the conclusions and discussion if the direction of the results was reversed.

2.6.1 Evidence of effectiveness

Medical systematic reviews assess the strength of inferences about the effectiveness of an intervention using guidelines that consider the strength of a causal inference (Hill 1971). Areas for consideration include:

1. The quality of the included studies
2. The size and significance of the observed effects
3. The consistency of the effects across studies or sites
4. The clarity of the relationship between the intensity of the intervention and the outcome
5. The existence of any indirect evidence that supports or refutes the inference
6. The lack of other plausible competing explanations of the observed effects (bias or confounding)

There are a range of approaches to grading the strength of evidence presented in medical reviews but there is no universal approach (Deeks *et al.* 2005). We suggest that authors of ecological reviews explicitly state weaknesses associated with each of the areas above, but the overall impact they make on conclusions can only be considered subjectively.

2.6.2 Applicability of results

End-users must decide, either implicitly or explicitly, how applicable the evidence presented in a systematic review is to their particular circumstances (Deeks *et al.* 2005). Authors should highlight where the evidence is likely to be applicable and equally importantly where it may not be applicable with reference to variation between studies and study characteristics.

Clearly, variation in the ecological context and geographical location of studies can limit the applicability of results. Authors should be aware of the timescale of included studies which may be insufficiently short to make long-term predictions. Variation in application of the intervention may also be important (and difficult to predict); but authors should be aware of differences between *ex situ* and *in situ* treatments (measuring efficacy versus effectiveness respectively) where they are combined, and should also consider the implications of applying the same intervention at different scales. Variation in baseline risk may also be an important consideration in determining the applicability of results as the net benefit of any intervention depends on the risk of adverse outcomes without intervention, as well as on the effectiveness of the intervention (Deeks *et al.* 2005). Given the myriad of factors involved in nature-conservation decision making, consideration of baseline risk is probably best left to end-users. However, reviewers should point out any clear discrepancies between high and low baseline risk groups where there is *a priori* rationale for the split.

Where reviewers identify predictable variation in the relative effect of the intervention in relation to the specified reasons for heterogeneity these should be highlighted. However, these relationships require cautious interpretation (because they are only correlations) particularly where sample sizes are small, data points are not fully independent, and where multiple confounding occurs.

Stage 3 - Reporting and dissemination of results

Wide dissemination and open access are key requirements of the evidence-based framework. For systematic reviews to have a real impact in terms of knowledge transfer from the science to the practitioner and policy communities they need to be readily accessible from a recognised central source. To this end the Collaboration for Environmental Evidence has established an independent, not-for profit, library of systematic reviews with the intent of managing and servicing the library in a similar format to the Cochrane Collaboration Library in medicine (see www.environmentalevidence.org) with its emphasis on transparency of the review process and independence from bias (Fazey et al. 2004). We urge reviewers to submit their reviews to the library and contact CEE at the earliest possible stage of the review process. The following sets out the principles and conditions for inclusion of reviews in the library.

Before reports are disseminated they should be subjected to expert scrutiny and peer review, including assessment of scientific quality and completeness. This process is organized by the CEE and is equivalent to that of a journal or grant board, but with a more supportive role in helping reviewers achieve the necessary quality rather than accepting/rejecting outright. If the CEE is contacted at an early stage and open consultation is undertaken as set out above then the chances of meeting the required standard should be significantly improved.

The full review should be submitted to CEE in the standard format (see Review Presentation and Formatting Guidelines at www.environmentalevidence.org). The format for reporting on the CEE website is a short summary that highlights the main review outcomes. This should be written so as to enable effective communication with managers and policy formers. A full review will normally include too much detail for wider dissemination but will nevertheless be made available, along with the summary, to all who want more information on the conduct of the review process. By mutual agreement, other formats such as policy briefs and guidance notes may also be posted.

Submitting and posting a review on the CEE website DOES NOT prevent further publication and the review may also be submitted, at the author's discretion, for publication in a peer-reviewed journal. All rights remain with the authors.

4.0 Requirement for further work

Systematic review in conservation and environmental management is in its infancy and these guidelines will need updating on a regular basis as well methodology develops from undertaking more reviews on a wider range of subjects. For example, most reviews to date have incorporated comparators, although work in progress involves synthesising experience and evidence employing Bayesian methodologies (Morris 1992; Louis 1993).

Other issues require consideration to strengthen the ecological guidelines presented above. Medical systematic review methodology is developing rapidly, with new techniques being developed to handle diverse types of variable quality data in fields such as diagnostic testing. The utility of these techniques for ecological purposes requires further investigation. Likewise, techniques for economic cost-benefit evaluation and disseminating evidence to different audiences (policy, scientific, practitioner and stakeholder groups) (NHMRC 2000) warrant consideration. Addressing all these issues is beyond the scope of these guidelines, but require further development if an ecological evidence base is to be fully established. The ecological guidelines presented evolved from the existing medical model. Table 3 highlights key differences between ecological and medical guidelines at present. As was the experience in the medical field, it will take time for systematic reviews to be recognized and valued as equivalent to other scientific papers in conservation. We hope these guidelines will set standards and facilitate key steps forward in encouraging more systematic reviews (e.g. journals encouraging their submission and publication and funders seeing systematic reviews as a valid form of research). We call on the conservation and environmental management community to engage with the CEE to further develop the library of systematic reviews and create the accessible evidence base that conservation and environmental management urgently requires.

Table 3. Differences between the medical systematic review guidelines and the ecological review guidelines advocated by the authors

Review stage	**Medical guidelines**	**Ecological guidelines**
Question formulation	Question formulation generally not limited by complexity and study numbers	Question formulation usually limited by information availability and complexity requiring a balance between holism (more realistic) and reductionism (more studies)
	Stakeholder engagement useful but not generally critical	Stakeholder engagement may be critical because conservation actions often result in conflicts in objectives
Developing review protocol: Search strategy	Complex searches balancing sensitivity and specificity are possible and recommended	High sensitivity, low specificity searches are recommended to reduce bias and increase repeatability because ecology lacks the sophisticated search infra-structure of medicine
Assessing quality of methodology	Clear hierarchy of evidence generally applicable and often used to define a minimum quality threshold	Pragmatic quality weightings and sensitivity analyses must augment data quality hierarchies to avoid misinterpretation, particularly when combining data across the hierarchy to increase sample sizes
	Performance bias and detection bias addressed by blinding and easy to assess using published quality weightings. Attrition bias common	Performance bias and detection bias addressed by experimental design and hard to assess especially in a standardized manner, necessitating the use of review specific quality weightings. Attrition bias rare
	Numerous "off the shelf" checklists to assess the validity of medical research	No "off the shelf" checklists hence the need for *a priori* review specific criteria preferably validated by consensus with stakeholders
Data extraction	Data extraction often relatively straightforward except for missing data and data hygiene problems	Data extraction complex especially with respect to variance measures for weighting. *A priori* rules must be developed in order to extract data in a repeatable standardized manner with independence and (pseudo)replication commonly problematic.
Data synthesis: Meta-analysis	Fixed and random effects models applicable	Random effects models generally more useful than fixed effect models because the complex interactions in ecology generally result in ecologically important heterogeneity between studies.

Acknowledgements
We thank the many conservation managers and scientists who have given us constructive feedback on the review process. Medical colleagues have contributed guidance particularly, T. Knight, R. Taylor, and K. Khan. We also thank members of the NCEAS meta-analysis working group and the Society for Research Synthesis Methodology for their interactions. This work was supported through grants from English Nature, the U.K. Natural Environment Research Council, the Royal Society for the Protection of Birds.

Literature Cited

Arnqvist, G., and D. Wooster 1995. Metaanalysis – synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution* **10**: 236–240.

Bero L., R. Grilli, J. Grimshaw, G. Mowatt, A. Oxman, and M. Zwarenstein eds. 1999. *Effective Practice and Organisation of Care Module of The Cochrane Database of Systematic Reviews. Issue 2. Update software.* The Cochrane Library, Oxford, United Kingdom.

Binns, N. A. and Remmick, R. 1994. Response of Bonneville Cutthroat Trout and Their Habitat to Drainage-Wide Habitat Management at Huff Creek, Wyoming. *North American Journal of Fisheries Management* **14**(4): 669-680**.**

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

Cooper, H.M. 1984. *Integrating Research. A Guide for Literature Reviews.* Sage Publications, Newbury Park.

Côté, I.M., Mosqueira, I. & Reynolds, J.D. 2001. Effects of marive reserve characteristics on the protection of fish populations: a meta-analysis. *Journal of Fish Biology* **59**: SA178-189

Deeks, J.J., J.P.T. Higgins and D.G. Altman 2005. "Analysing and presenting results". In: *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5* [updated May 2005]; Section 8. (ed by J.P.T. Higgins and S. Green.) www.cochrane.org/resources/handbook/hbook.htm (accessed 12th July 2006).

Dixon-Woods, M., S. Agarwal, B. Young. D. Jones, and A. Sutton. 2004. *Integrative approaches to qualitative and quantitative evidence*. National Health Services Health Development Agency, London.

Dogpile. 2007. *Different Engines, Different Results: Web Searchers Not Always Finding What They're Looking for Online*. Available at: www.infospaceinc.com/onlineprod/OverlapDifferentEnginesDifferentResults.pdf Accessed: 31/03/2008.

Downing, J.A., Osenberg, C.W. & Sarnelle, O. 1999. Meta-analysis of marine nutrient-enrichment experiements: variation in the magnitude of nutrient limitation. *Ecology* **80**: 1157-1167.

Edwards, P., M. Clarke, C. DiGuiseppi, S. Pratap, I. Roberts, and R. Wentz. 2002. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine* **21**: 1635-1640.

Emerson, J.D., E. Burdick, D.C. Hoaglin, F. Mosteller, and T.C. Chalmers. 1990. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials* **11**: 339-352.

Eysenbach G., Tuische J., and Diepgen T. L. 2001. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. *Medical Informatics and the Internet in Medicine*, 26(3):203-218.

Fazey, I., J.G. Salisbury, D.B. Lindenmayer, J. Maindonald, and R. Douglas. 2004. Can methods applied in medicine be used to summarize and disseminate conservation research? *Environmental Conservation* **31**: 190-198.
Feinstein, A.R. 1985. *Clinical Epidemiology: The Architecture of Clinical Research.* Saunders, Philadelphia.
Feinstein, A.R., and R.I. Horwitz. 1982. Double standards, scientific methods, and epidemiological research. *New England Journal of Medicine* **307**: 1611-1617.
Fox, D.M. 2005. Evidence of evidence-based health policy: The politics of systematic reviews in coverage decisions. *Health Affairs* **24**: 114-122.
Gates, S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology* **71**: 547–557.
Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 1995 Bayesian data analysis. *Texts in Statistical Science.* Chapman & Hall, London, UK.
Gotzsche, P. C. 1987. Reference bias in reports of drug trials. *British Medical Journal* **295**: 654-656.
Gurevitch, J. and L.V. Hedges 2001. *Meta-analysis Combining the results of independent experiments.* In: “Design and Analysis of Ecological Experiments” (ed by S.M. Scheiner and J. Gurevitch) pp. 347-369. Oxford University Press, New York.
Hedges, L.V. 1994. “Statistical considerations”. In: *The Handbook of Research Synthesis.* (ed by H. Cooper and L.V. Hedges) pp. 30-33. Russell Sage Foundation, New York.
Higgins, J.P.T. and S. Green. (eds) 2005. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5.* John Wiley & Sons, Ltd, Chichester, UK.
Hill, A.B. 1971 Principles of Medical Statistics. *Lancet* **9**: 312-20.
Hock, R. 1999. *The Extreme Searcher’s Guide to Web Search Engines: A Handbook for the Serious Searcher*. CyberAge Books, New Jersey, USA.
Horwitz, R.I., and A.R. Feinstein. 1979. Methodological standards and contradictory results in case-control research. *American Journal of Medicine* **66**: 556-564.
Introna, L. D. and Nissenbaum, H. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16: 169-185.
Jackson, G.B. 1980. Methods for integrative reviews. *Review Education Research* **50**: 438-460.
Jüni, P., A. Witschi, R. Bloch, and M. Egger. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of American Medical Association* **282**: 1054-60.
Khan, K.S., R. Kunz, J. Kleijnen and G. Antes. 2003. *Systematic reviews to support evidence-based medicine: how to apply findings of healthcare research.* Royal Society of Medicine Press Ltd, London.
Kunz, R., and A.D. Oxman. 1998. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised trials. *British Medical Journal* **317**: 1185-1190.
Lawrence, S. and Giles, C. L. 1998. Searching the World Wide Web. *Science*, 280: 98-100.
Leimu, R., and J. Koricheva 2005. What determines the citation frequency of ecological papers? *Trends in Ecology and Evolution* **20:** 28-32.
Levine, M., S. Walter, H. Lee, T. Haines, A. Holbrook, and V. Moyer. 1994. The Evidence-Based Medicine Working Group. Users' guides to the medical literature IV: how to use an article about harm. *Journal of American Medical Association* **271**:1615-1619.

Light R.J., and D.B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Harvard University Press, Massachusetts.

Lipsey, M.W. and D.B. Wilson 2001. Practical Meta-analysis. *Applied Social Research Methods Series*. Volume 49. Sage Publications, Thousand Oaks, California.

Louis. T., and D. Zelterman. 1993. "Bayesian approaches to research synthesis". In: *The Handbook of Research Synthesis*. (ed. by H. Cooper and L.V. Hedges), Pp 411-422. Russell Sage Foundation, New York.

Moher, D., A.R. Jadad, G. Nichol, M. Penman, P. Tugwell, and S. Walsh. 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials* **16**: 62-73.

Moher, D., A.R. Jadad, and P. Tugwell. 1996. Assessing the quality of randomized controlled trials: current issues and future directions. *International Journal of Technology Assessment in Health Care* **12**: 195-208.

Morris, C.N., and S.L. Normand. 1992. "Hierarchical models for combining information and for meta-analyses". In: *Bayesian Statistics* (ed. By J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), pp. 321-344. 4th Ed. Oxford University Press, New York.

NHS Centre for Reviews and Dissemination. 2001. *Undertaking systematic review of research on effectiveness*. NHS CRD, University of York.

Osenberg, C.W., O. Sarnelle, S.D. Cooper and R.D. Holt. 1999. Resolving ecological questions through meta-analysis: goals, metrics and models. *Ecology* **80**: 1105–1117.

Petticrew,M. and Roberts, H. 2006. *Systematic Reviews in the Social Sciences. A practical guide.* Blackwell, UK

Pullin, A., and T. Knight. 2001. Effectiveness in conservation practice: pointers from medicine and public health. *Conservation Biology* **15**: 50-54.

Pullin, A., and T. Knight. 2003. Support for decision-making in conservation practice: an evidence-based approach. *Journal for Nature Conservation* **11**: 83-90.

Pullin, A., T. Knight, D. Stone, and K. Charman. 2004. Do conservation managers use scientific evidence to support their decision-making? *Biological Conservation* **119**: 245-252

Ravnskov, U. 1992. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *British Medical Journal* **305**: 9-15**.**

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.

Sharp, S. 1998. Meta-analysis regression: statistics, biostatistics, and epidemiology. *Stata Technical Bulletin* **42**: 16-22.

Schulz, K.F., I. Chalmers, R.J. Hayes, and D.G. Altman. 1995. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* **273**: 408-412.

Schafer, J.L. (1997) Analysis of incomplete multivariate data. *Monographs on Statistics and applied Probability* **72**. Chapman & Hall, Boca Raton USA.

Smart, J. M. and Burling, D. 2001. Radiology and the Internet: A Systematic Review of Patient Information Resources. *Clinical Radiology*, **56**: 867-870.

Stevens, A., and R. Milne. 1997. "The effectiveness revolution and public health". In: *Progress in Public Health* (ed. By G. Scally), Pp. 197-225. Royal Society of Medicine Press, London.

Stewart, G.B., Bayliss, H.R., Showler, D.A., Pullin, A.S., and Sutherland, W.J. 2006. Does the use of in-stream structures and woody debris increase the abundance of salmonids? *CEE Systematic Review 12.* Available from http://www.environmentalevidence.org/ CEELibraryofsystematicreviews.htm

Stewart, G. B., C. F. Coles, and A. S. Pullin. 2005. Applying evidence-based practice in conservation management: lessons from the first systematic review and dissemination projects. *Biological Conservation* **126**: 270-278.

Stewart, G.B., A.S. Pullin, and C.F. Coles. 2005. Effects of wind turbines on bird abundance. *CEE Systematic Review 4.* Available from www.environmentalevidence.org/CEELibraryofsystematicreviews.htm

Sutherland, W., A. Pullin, P. Dolman, and T. Knight. 2004. The need for evidence-based conservation. *Trends in Ecology and Evolution* **19**: 305-308.

Thompson, S. 1994. Systematic review: why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* **309:** 1351-1355.

Tyler, C., E. Clark, and A. S. Pullin. 2005. Do management interventions effectively reduce or eradicate populations of the American Mink, Mustela vison? *CEE Systematic Review 7.* Available from www.environmentalevidence.org/CEELibraryofsystematicreviews.htm

Tyler, C., and A.S. Pullin. 2004. Do commonly used interventions effectively control *Rhododendron ponticum*? *CEE Systematic Review 6.* Available from www.environmentalevidence.org/CEELibraryofsystematicreviews.htm

Tyler, C., A.S. Pullin, and G. B. Stewart. 2006. Effectiveness of management interventions to control invasion by *Rhododendron ponticum. Environmental Management* **37**, 513- 522.

University of California 2008. *Finding information on the Internet: a Tutorial - Meta-Search Engines.* Available at: http://www.lib.berkeley.edu/TeachingLib /Guides/Internet/MetaSearch.html Accessed 31/03/2008.

Appendix A

Example – Scoping: the iterative development of a database search strategy

The below example is based on pre-review scoping conducted for Systematic Review 48, "The Evidence Base for Community Forest Management as a Mechanism for Supplying Environmental Benefits and Improving Rural Welfare" (see http://www.environmentalevidence.org/SR48.html) and is presented as an illustration of the iterative nature of search term development.

Scoping searches were conducted in Web of Knowledge with the objective of testing the utility of the stakeholder-suggested search terms (see Table 1 below) and providing an idea of the potential numbers of returned hits to guide resource planning. The suggested search terms were split into three groups: the first based on the intervention of interest, the second guided by the outcome elements of the review question, and the third influenced by the types of study of interest (Table 1). Only if searches based on set one returned an unmanageable number of hits would it have been appropriate to use sets two and three.

Table 1. Original stakeholder-proposed search terms.

Set:	**Search terms:**
One	Community Forest Management Co-management forest Joint management forest Participatory management forest Indigenous forest reserve Decentralized Forest Governance Community engagement in forest management
Two	Biodiversity, desert*, degrad*, economic, carbon, poverty, fuel*
Three	evidence, empirical, quantitative, evaluation, assessment, measures

The results shown below in Table 2 illustrate the evolution of this set of terms, from one returning a huge number of spurious hits, to one more sensitive and manageable. On the basis of these findings, it was thus deemed appropriate to exclude the terms suggested in sets two and three, as it was felt that these may have been overly restrictive in this context.

Table 2. Search term scoping and evolution.

Search string	**Number of hits (Web of Knowledge)**	**Change from previous**
1. Topic=((community forest management) OR (co-management forest*) OR (joint management forest*) OR (participatory forest*) OR (indigenous forest* reserve*) OR (decentrali* forest*) OR (integrated conservation development pro*) OR (ICDP*))	21,464	n/a
2. Topic=("community forest management" OR "co-management		

forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	250	Quotation marks added to improve % relevance
3. Topic=("community forest* management" OR "co-management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	256	Wildcard added to pick up alternative word endings in first phrase
4. Topic=("community forest* management" OR "co-management forest*" OR "co management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	256	De-hyphenated variant added for co-management phrase. Not useful
5. Topic=("community forest*" OR "co-management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,008	1st phrase amended ("management" removed) to pick up alternatives such as "community forestry" or "community forests", etc.
6. Topic=("community forest*" OR "forest* co-management " OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,019	2nd phrase amended to more probable word order
7. Topic=("community forest*" OR "forest* co-management " OR ("joint management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,035	Third phrase amended to pick up all variants of the term – e.g. "forest joint management" or "joint management forests/ry, etc."
8. Topic=("community forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,096	Ditto above for second phrase
9. Topic=("community forest*" OR ("co-management " AND forest*) OR ("joint management" AND		

forest*) OR "JFM" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,264	"JFM" noted as a standalone term in some of the Indian literature, and thus included
10. Topic=("community forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "JFM" OR "participatory forest*" OR ("collaborative management" AND forest*) OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,279	Addition of further 'intervention' term
11. Topic=("community forest*" OR "community-based forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "JFM" OR "participatory forest*" OR ("collaborative management" AND forest*) OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,304	Ditto above
12. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" AND "social forestry")	15,195	Addition of 'social forestry'. Deemed too broad to be useful. Nothing apparently additional retrieved.
13. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource")	1385	"Community based natural resource" added – apparently very useful
14. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated	1563	(community AND "natural resource management" AND forest*) added to account

conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*))		for alternative variants
15. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*) OR "common property")	3344	"Common property" added but broad
16. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*) OR ("common property" AND forest*))	1715	"forest*" added to common property phrase to restrict spurious hits
17. Topic=("community management" AND woodland*)	13	Not useful – all relevant papers either contained term 'forest' or other 'intervention' based terms e.g community-based natural resource management
18. Topic=("community management" AND tree*)	39	Ditto above
19. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management"	1715	SUGGESTED TERMS (FOR DRAFT PROTOCOL)

AND forest*) OR ("common property" AND forest*))		

* indicate the use of wildcards or ‘truncation’, to search for variant word endings. Terms in red font are those omitted or included at each stage.