Ecological Applications, 26(5), 2016, pp. 1295–1301

An evidence assessment tool for ecosystem services and conservation studies

Anne-Christine Mupepele,[1,3] Jessica C. Walsh,[2] William J. Sutherland,[2] and Carsten F. Dormann[1]

[1]*Department of Biometry and Environmental System Analysis, University of Freiburg, Tennenbacherstr. 4, 79106 Freiburg, Germany*
[2]*Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, United Kingdom*

Abstract. Reliability of scientific findings is important, especially if they directly impact decision making, such as in environmental management. In the 1990s, assessments of reliability in the medical field resulted in the development of evidence-based practice. Ten years later, evidence-based practice was translated into conservation, but so far no guidelines exist on how to assess the evidence of individual studies. Assessing the evidence of individual studies is essential to appropriately identify and synthesize the confidence in research findings. We develop a tool to assess the strength of evidence of ecosystem services and conservation studies. This tool consists of (1) a hierarchy of evidence, based on the experimental design of studies and (2) a critical-appraisal checklist that identifies the quality of research implementation. The application is illustrated with 13 examples and we suggest further steps to move towards more evidence-based environmental management.

Key words: *governance; quality checklist; quantification; rigour; valuation.*

Introduction

Conservation and ecosystem services studies are important scientific sources for decision-makers seeking advice on environmental management (Daily and Matson 2008, Kareiva and Marvier 2012). Their results potentially influence actions and it is therefore crucial to assess transparently the reliability of current research and its recommendations (Pullin and Knight 2003, Boyd 2013).

Evidence-based practice was introduced in the medical field aiming to assess the reliability of scientific statements and identify the best available information to answer a question of interest (Sackett et al. 1996, GRADE Working Group 2004, OCEBM Levels of Evidence Working Group 2011). In conservation, evidence-based practice was first mentioned 15 yr ago (Sutherland 2000, Pullin and Knight 2001). Today, the Collaboration for Environmental Evidence fosters the creation of systematic reviews to collate the strongest possible evidence (Petrokofsky et al. 2011, Collaboration for Environmental Evidence 2013; see also Journal for Environmental Evidence), together with Conservation Evidence (Hopkins et al. 2015), which focuses on the development of summaries and guidelines, and the communication of evidence to practitioners (Sutherland et al. 2012, Dicks et al. 2014). Summaries, contrary to systematic reviews, do not focus on a specific question but bring together information from a much broader topic, e.g., from a whole animal group, such as bees (Dicks , et al. 2010, 2014, Walsh et al. 2015).

Systematic reviews and summaries compile individual studies and therefore require the evaluation of the evidence at the level of the individual study. In systematic reviews this is typically mentioned as one step of the critical appraisal. However, to date, such critical appraisal is often implicit, based on criteria varying for every systematic review (Collaboration for Environmental Evidence 2013, Carroll and Booth 2015, Stewart and Schmid 2015). We therefore introduce an evidence assessment tool providing a clear appraisal guideline to score the reliability of individual studies.

Definitions and Terminology

A well-defined terminology is essential for effective communication between practitioners and scientists. Evidence is the "ground for belief" or "the available body of information indicating whether a belief or proposition is true or valid" (Howick 2011). Evidence describes the knowledge behind a statement and expresses how solid our recommendations are (see also Higgs and Jones 2000:311; Rychetnik et al. 2001, Lohr 2004, Binkley and Menyailo 2005, Pullin and Knight 2005). The strength of evidence reflects the reliability of information and we can identify whether a statement is based on strong or weak evidence, i.e., very reliable or hardly reliable. Hence evidence-based practice means to identify the reliability of current knowledge, based on research integrated with

Manuscript received 14 April 2015; revised 6 November 2015; accepted 23 November 2015. Corresponding Editor: D. Schimel.
[3]E-mail: anne-christine.mupepele@biom.uni-freiburg.de

expertise, and to act according to this best available knowledge. The collation and appraisal of the best available evidence follow strict criteria to ensure transparency and to reduce bias. A goal of evidence-based practice is to act on best available evidence while being aware of the strength of inference this evidence permits (Howick 2011:15).

Setting Question and Context

The formulation of a clear research question and the purpose of investigation is highly emphasized throughout the evidence literature (Higgins and Green 2011, Collaboration for Environmental Evidence 2013:20–23). Questions should specify which ecosystem service, species or aspect of biodiversity will be investigated in which system, as this will help to determine the external validity of the answer provided in a study.

We further recommend to determine the focus of the question as either quantification, valuation, management, or governance. Quantification studies measure the amount of an ecosystem service, species abundance, biodiversity, or other conservation targets. Measures can be taken in absolute units or relative to another system. Valuation studies assess the societal value of ecosystem services. The most common way is monetary valuation. Management is the treatment designed to improve or benefit specific ecosystem services, target species, or other conservation aspects. For example, leaving dead wood in forests to increase biodiversity or reducing agricultural fertilizer to decrease nearby lake eutrophication. Governance is seen as the strategy or policy to steer a management intervention, such as REDD (Reducing Emissions from Deforestation and Forest Degradation), which aims to encourage forest protection and reforestation (Kenward et al. 2011). The strategies used by policy makers include incentives (subsidiaries) or penalties (law/tax; see also Bevir 2012). When the effectiveness of management and governance strategies is determined, evidence-based quantification or valuation is required to measure the outcome of the management or governance intervention. Acuña et al. (2013), for example, used valuation methods to determine success or failure of a management strategy, while Walsh et al. (2012) quantified malleefowl abundance through monitoring survey data to assess the management impact of fox baiting. The distinction of four different foci is essential to assess the whole range of environmental management.

We have described how to set the context of questions that can be useful in environmental management. Once the question has been determined, and the investigation carried out, the strength of the resulting evidence should be assessed (Fig. 1).

Evidence Assessment

The reliability of a study is characterized by its study design and the quality of its implementation. Both are evaluated in the evidence assessment.

Evidence hierarchy

The study design refers to the set-up of the investigation, e.g., controlled or observational design (GRADE Working Group 2004). These study designs are not equally compelling with respect to inferring causality. Differences in study designs typically translate into weak or strong evidence. To identify the reliability of a study, study designs can be ranked hierarchically according to a level-of-evidence scale, henceforth, the evidence hierarchy (Fig. 2).

Systematic reviews (LoE1a) are at the top of the evidence hierarchy (Fig. 2) and provide the most reliable

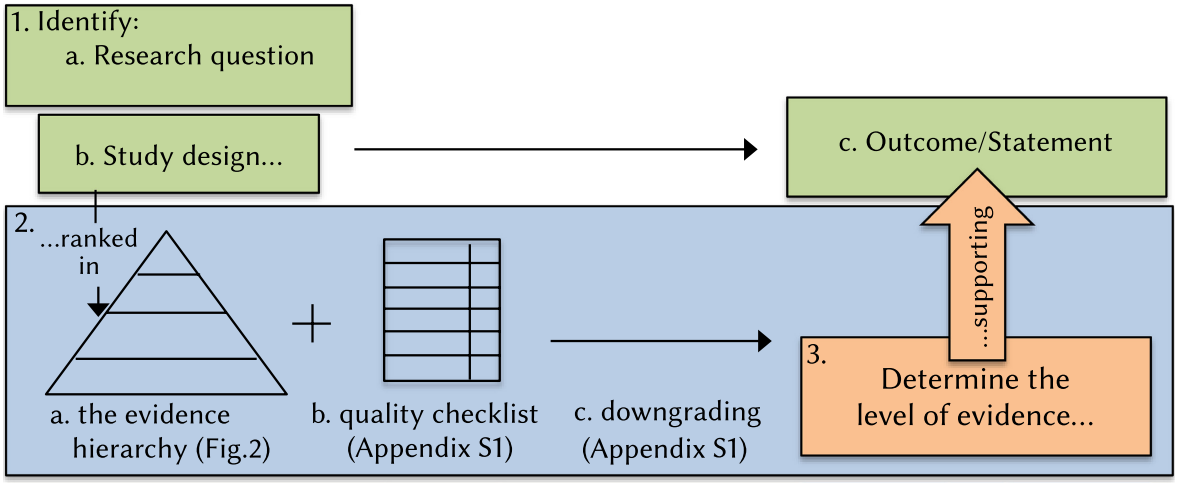

Fig. 1. Schematic representation of the evidence assessment tool. (1) Identification of study question, design, and outcome. (2) Assessing a level of evidence based on the underlying study design and calculating a quality score based on the quality checklist. (3) Determining the final level of evidence supporting the outcome by downgrading the originally assigned level of evidence according to the quality score.

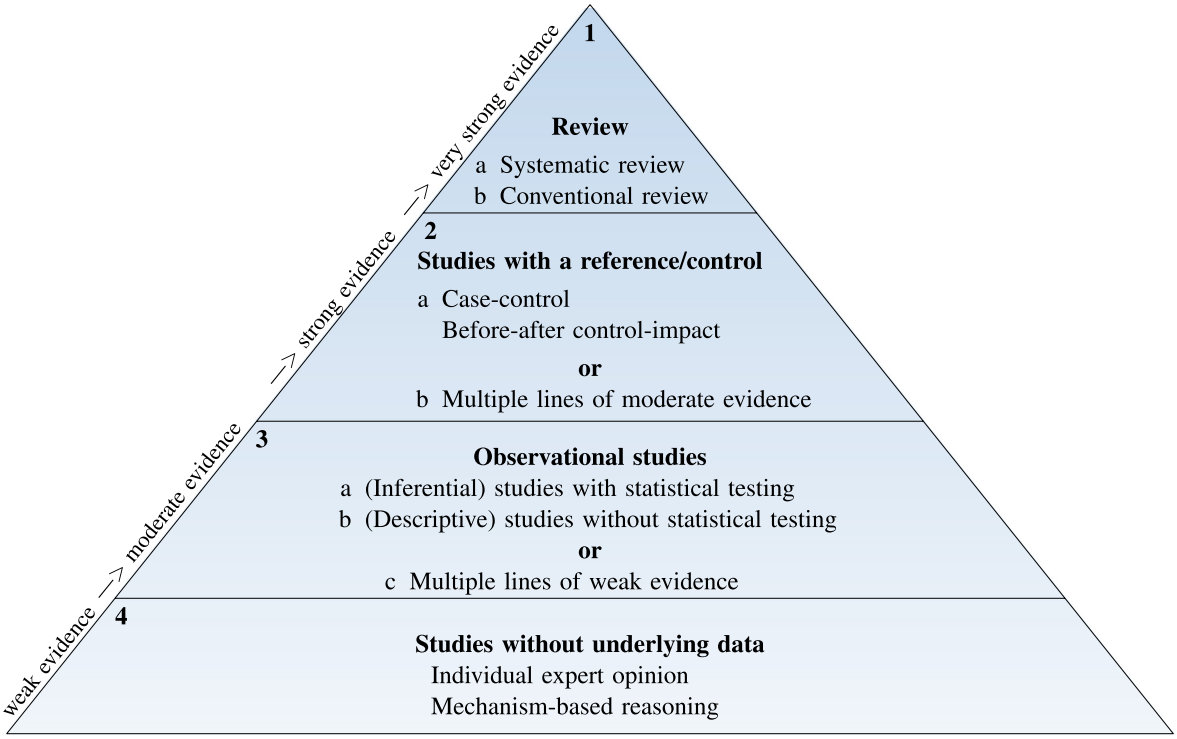

FIG. 2. Level-of-evidence (LoE) hierarchy ranking study designs according to their evidence. Very strong evidence (LoE1) to weak evidence (LoE4) with internally ranked sublevels a, b, and c.

information. They summarize all information collated in several individual studies, have an a priori protocol on design and procedure, and are conducted according to strict guidelines (e.g., Collaboration for Environmental Evidence 2013). If possible, they ideally include quantitative measures, i.e., a meta-analysis (see Koricheva et al. 2013, Vetter et al. 2013). All other, non-systematic and more conventional reviews (LoE1b) may also include quantitative analysis or are purely qualitative. Both types of review summarize the findings of several studies, but systematic reviews assess the completeness and reproducibility more carefully and strive to reduce bias by having transparent, thorough, pre-defined methods (Freeman et al. 2006, Higgins and Green 2011, Collaboration for Environmental Evidence 2013, Haddaway and Bayliss 2015, Haddaway and Bilotta 2015).

The necessary condition for any review is that appropriate individual studies are available. The most reliable individual study design is a study with a reference/control (LoE2). Typically, these are case-control or before–after control–impact studies (LoE2a; Smith et al. 2014). Investigations that cannot follow such a controlled design may alternatively seek to gain strong evidence through multiple lines of moderate evidence (LoE2b). Multiple lines of evidence require at least two unrelated and consistent arguments to confirm the study conclusions, thereby forming a non-contradicting picture (see also Smith et al. 2002). Illustrative examples are the valuation of ecosystem services (e.g., Mogas et al. 2006) or long-term environmental processes that are difficult to control (e.g., Dorman et al. 2015). Multiple lines of evidence can be collected in individual studies using different approaches within one study context (LoE2b, LoE3c) or in reviews (LoE1) including evidence from different studies.

Observational studies (LoE3) are individual studies without a control. These include studies employing inferential and correlative statistics (LoE3a), e.g., testing for the influence of environmental variables on the quantity of an ecosystem service. Descriptive studies (LoE3b) imply data collection and representation without statistical testing (e.g., data summaries, ordinations, histograms, surveys). Multiple lines of weak evidence (LoE3c) can increase the evidence of LoE4 investigations; elicitation of independent expert opinions is a well-known example (Sutherland et al. 2013, Morgan 2014, Smith et al. 2015, Sutherland and Burgman 2015; see also Appendix S1).

The lowest level of evidence are statements without underlying data (LoE4). These are usually individual expert opinions, often not distinguishable from randomness (Tetlock 2005, Drolet et al. 2015). Other statements without underlying data are reasoning based on mechanism. Mechanism-based reasoning involves an inferential chain linking an intervention to the outcome (Howick et al. 2010, Howick 2011). If this chain of mechanisms is not supported by data, there is no possibility to assess whether all relevant mechanisms linking the

intervention to the outcome have been included. Mechanism-based reasoning without corroborative data provides only weak evidence. On the other hand, mechanism-based reasoning can result in a model that is validated and tested on real world data. With such a data validation, the model could reach moderate evidence or strong evidence, depending on the underlying study design.

It is important to note that method and design should not be confused. Methods are the means used to collect or analyze data, e.g., remote sensing, questionnaires, or ordination techniques. Design reflects how the study was planned and conducted, e.g., a case-control or observational design (GRADE Working Group 2004). The same methods can be employed for different underlying designs. Remote sensing, for example, can be done purely descriptively (LoE3b) or with a reference such as ground-truthing or in a before-and-after design (LoE2a). Analogously, models can represent theories without supporting data (LoE4), involve data input to determine parameters (LoE3b), or be tested and validated (LoE3a). To achieve strong evidence, model predictions have to be confirmed by several unrelated data sets forming a non-contradicting picture (LoE2b) or should be built on information derived from controlled studies unequivocally identifying the underlying causal mechanism (LoE2a; Kirchner 2006).

Critical appraisal

Study design alone is an inadequate marker of the strength of evidence (Rychetnik et al. 2001). A study with a strong-evidence design may be poorly conducted. The critical appraisal assesses the implementation of the study design, specifically the methodological quality, the actual realization of the study design, and its reporting (Higgins and Green 2011). It identifies the study quality and may lead to a downgrading in the evidence hierarchy. Quality, in this context, is the extent to which all aspects of conducting a study can be shown to protect against bias and inferential error (Lohr 2004). Quality checklists can be used to detect bias and inferential error. Combining 30 published quality checklists, we provide the first quality checklist for conservation and ecosystem services (Appendix S1: Table S1), that can be used to comprehensively assess the internal validity of a study, covering questions on data collection, analysis, and the presentation of results. The checklist consists of 43 questions, of which some apply only to a specific context, e.g., for reviews or studies focusing on valuation. All questions answered with yes receive one point. In the case of non-reported issues, we advise the answer no to indicate a deficient reporting quality. The percentage of points received can help to decide whether to downgrade the level of evidence (Appendix S1: Table S2).

Reviews provide information at the highest level of evidence and their critical appraisal is different from other designs because they are based on studies with weaker evidence (see Appendix S1: Table S1, Review). Every single study included in the review can be assessed for its level of evidence using the evidence hierarchy and the checklist for quality criteria. If only studies based on weak evidence were included, then the review should be downgraded, regardless of other quality criteria. In addition, a review can be assessed for other quality shortcomings using again the quality checklist.

The checklist should make the assessment more transparent, but we are aware that the process may not always be straightforward. Questions in the checklist can be subjective and depend on the judgment of the assessor. Cohen's kappa test was used to test the agreement in 13 exemplary studies between two different assessors (Appendix S1: Table S3). It ranges from 0 to 1, representing random to perfect agreement. Our result revealed a moderate agreement (unweighted Cohen's kappa = 0.49; P-value <0.001; Cohen 1960, Landis and Koch 1977, Gamer et al. 2015). Depending on the context, the assessor may decide to give more weight to particular questions or add questions to the checklist. Although the procedure cannot be fully standardized, we are not aware of a better alternative, and we encourage the use of the checklist as a baseline that can be adapted for specific studies.

The combination of study design (Fig. 2) and quality criteria (Appendix S1: Table S1) is the last step and identifies the strength of evidence supporting the study result (schematic representation in Fig. 1). The level of evidence derived by the study design should be downgraded depending on the quality score calculated from the quality checklist (Appendix S1: Table S2).

Application of the Evidence Assessment Tool

The suggested method was applied to assess the evidence of 13 studies (Appendix S1: Table S3). They were selected to serve as examples and illustrate the applicability of the tool to the whole range of study designs and foci. The first example was a management-related systematic review of Mant et al. (2013), conducted according to the guidelines of the Collaboration for Environmental Evidence (2013). They investigated the effect of liming rivers or lakes on fish and invertebrate populations. They found that liming increased fish abundances and acid-sensitive invertebrates, but may have a negative impact on the abundance of all invertebrate taxa combined. According to the critical appraisal, the study achieved 21 out of 24 points (88%) and it therefore remained at the originally assigned LoE1a, the highest level of evidence (Appendix S1: Table S3).

A second example tackles the question: How does adding dead wood to rivers influence the provision of ecosystem services? (Acuña et al. 2013). The authors investigated two ecosystem services (fishing and retention of organic and inorganic matter) in a river-forest ecosystem in Spain and Portugal and studied the effect of this management intervention. Their study design followed a before–after control–impact approach, equivalent to LoE2a. The critical appraisal revealed shortcomings, e.g., no blinding, no randomization, and no probability sampling: only 17 out of 25 points (68%) were

achieved. The level of evidence was downgraded by one level to LoE3a. We therefore conclude that the statement made by Acuña et al. (2013): "restoration of natural wood loading in streams increases the ecosystem service provision" is based on moderate evidence (LoE3a).

We provide further examples in the Appendix (Appendix S1: Tables S3 and S4). All but one study revealed quality shortcomings and had to be downgraded. Most were scored as LoE3 or LoE4.

Relevance for Different User Groups

In the previous section it was elaborated how to assess the strength of evidence for individual studies and reviews. Now we provide a few notes on *who* should use the evidence assessment tool.

1. Scientists conducting their own studies have to be aware of how to achieve strong evidence, particularly during the planning phase. Choosing a study design that provides strong evidence and respects the quality criteria will substantially increase the potential contribution to our knowledge.
2. Scientists advising decision-makers should be explicit about the strength of evidence of information they include in their recommendations. Weighting all scientific information equally, or subjectively, runs the risk of overconfidence and bias.
3. Decision-makers receiving information from scientists should demand a level-of-evidence statement for the information provided. Alternatively, they can assess the strength of evidence themselves. However, this may be difficult as it takes time and requires some scientific training to identify the study design and evaluate the quality questions.
4. We further encourage consortia, international panels and learned societies, such as the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES), the Ecological Societies (ESA, BES, GFÖ, and others), the Society for Conservation Biology (SCB), and the Ecosystem Services Partnership (ESP) to support the development of guidelines that include an evidence assessment (Graham et al. 2011, Sutherland et al. 2015). These best-practice guides are based on the collection of scientific evidence synthesized and judged by a group of experts. They provide recommendations on how to best quantify, value, manage, or govern a desired ecosystem service or conservation target, giving decision-makers transparent advice with an emphasis on the strength of the evidence available (Graham et al. 2011).

Discussion

We have outlined an evidence assessment tool for ecosystem services and conservation studies, encompassing a hierarchy to judge the available evidence based on study design and a quality checklist to facilitate critical appraisal. We have further illustrated with examples how to apply the tool (see also Appendix S1: Tables S3 and S4).

Evidence-based practice seeks to complement existing management frameworks by emphasizing the importance of systematically collating the existing scientific evidence and assessing it for its reliability and relevance. The IPCC report, for example, uses a combined measure of evidence and level of agreement (Mastrandrea et al. 2010, Spiegelhalter and Riesch 2011). Our suggested approach is more detailed, describing how one can actually assess the evidence.

Evidence-based practice has faced criticism of its evidence hierarchies, claiming that controlled trials are not always more reliable than observational studies. A main argument against hierarchies is that they are rigid and only consider the study design to assign a level of evidence (Petticrew and Roberts 2003, Adams and Sandbrook 2013, Stegenga 2014). With our quality checklist, we emphasize the critical appraisal to check for an appropriate implementation and methodological quality of study designs. The proposed assessment therefore does not overestimate the results of deficiently implemented meta-analyses and controlled studies. Some science sectors have to rely on observational studies because their study units cannot be controlled. This usually applies to environmental governance, conservation biology of rare species, or global theories that lack a second earth as a control. Multiple lines of evidence can lead to strong evidence using only observational study designs (Fig. 2, LoE2b). However, a central task of natural science is to determine causal relationships, and observational studies do not have the same strength to determine causal relationships as replicated and randomized case-control studies (Holland 1986, Grimes and Schulz 2002, Illari et al. 2011). We should acknowledge that in some areas of science causality cannot be established, and hence the reliability achieved remains lower than in areas where it can.

Other criticism has been directed toward the fact that every system is unique and the external validity of studies is low. We are aware that generalizability of results is problematic in ecosystems, where many different drivers take influence at the same time and hence, the general evidence may not apply due to particular circumstances. At this point the judgment of experts on the external validity of the currently best available evidence is irreplacable (Karanicolas et al. 2008, Howick 2011). Evidence-based practice means integrating individual expertise with the best available evidence from systematic research (Sackett et al. 1996, Straus et al. 2010). More reflection and responses to criticism of evidence-based practice can be found in Mullen and Streiner (2004), Sutherland et al. (2004, 2005), and Haddaway and Pullin (2013).

Despite the criticism raised against evidence-based practice the benefits are clear (Gilbert et al. 2005, Howick 2011, Walsh et al. 2015). Rating the strength of evidence matters as it clarifies the reliability of research results and,

thus, the strength of conclusions, decisions, or recommendations drawn from that research (Lohr 2004).

Reliable scientific evidence in environmental management is pivotal, and its use (or misuse) can have immense impacts on environmental outcomes and the society. It is essential that scientists and decision makers consider the strength of evidence when conducting studies, provding advice, and taking decisions. In the interest of responsible use of environmental resources and processes, we strongly encourage embracing evidence-based practice as a paradigm for all research contributing to environmental management.

Acknowledgments

We thank Andrew Pullin, Sven Lautenbach, and Ian Bateman for valuable comments on earlier versions of the manuscript. This work was supported by the 7th framework programme of the European Commission in the project OPERAs (grant number 308393, www.operas-project.eu).

Literature Cited

Acuna, V., J. R. Diez, L. Flores, M. Meleason, and A. Elosegi. 2013. Does it make economic sense to restore rivers for their ecosystem services? Journal of Applied Ecology 50:988–997.

Adams, W. M., and C. Sandbrook. 2013. Conservation, evidence and policy. Oryx 47:329–335.

Bevir, M. 2012. Governance: a very short introduction. Oxford University Press, Oxford, UK.

Binkley, D., and O. Menyailo. 2005. Gaining insights on the effects of tree species on soils. Pages 1–16 *in* editor. Tree species effects on soils: implications for global change, chapter 1. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Boyd, I. 2013. A standard for policy-relevant science. Nature 501:159–160.

Carroll, C., and A. Booth. 2015. Quality assessment of qualitative evidence for systematic review and synthesis: is it meaningful, and if so, how should it be performed? Research Synthesis Methods 6:149–154.

Cohen, J. 1960. A coefficient of agreement for nomial scales. Educational and Psychological Measurement 20:37–46.

Collaboration for Environmental Evidence. 2013. Guidelines for Systematic Review and Evidence Synthesis in Environmental Management. Version 4.2. Environmental Evidence. URL www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf.

Daily, G. C., and P. A. Matson. 2008. Ecosystem services: from theory to implementation. Proceedings of the National Academy of Sciences 105:9455–9456.

Dicks, L. V., D. A. Showler, and W. J. Sutherland. 2010. Bee conservation: evidence for the effects of interventions. Pelagic Publishing, Exeter, UK.

Dicks, L. V., J. C. Walsh, and W. J. Sutherland. 2014. Organising evidence for environmental management decisions: a '4S' hierarchy. Trends in Ecology & Evolution 29:1–7.

Dorman, M., T. Svoray, A. Perevolotsky, Y. Moshe, and D. Sarris. 2015. What determines tree mortality in dry environments? a multi-perspective approach. Ecological Applications 25:1054–1071.

Drolet, D., A. Locke, M. A. Lewis, and J. Davidson. 2015. Evidence-based tool surpasses expert opinion in predicting probability of eradication of aquatic nonindigenous species. Ecological Applications 25:441–450.

Freeman, S. R., H. C. Williams, and R. P. Dellavalle. 2006. The increasing importance of systematic reviews in clinical dermatology research and publication. Journal of Investigative Dermatology 126:2357–2360.

Gamer, M., J. Lemon, I. Fellows, and P. Singh. 2015. irr: various coefficients of interrater reliability and agreement. R-package version 0.84 https://cran.r-project.org/web/packages/irr/irr.pdf.

Gilbert, R., G. Salanti, M. Harden, and S. See. 2005. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. International Journal of Epidemiology 34:874–887.

GRADE Working Group. 2004. Grading quality of evidence and strength of recommendations. BMJ 328:1–8.

Graham, R., M. Mancher, D. M. Wolmann, S. Greenfield, and E. Steinberg. 2011. Clinical practice guidelines we can trust. The National Academies Press, Washington, D.C., USA.

Grimes, D. A., and K. F. Schulz. 2002. Descriptive studies: what they can and cannot do. Lancet 359:145–149.

Haddaway, N. R., and H. R. Bayliss. 2015. Clarification on the applicability of systematic reviews. Frontiers in Ecology and the Environment 13:129.

Haddaway, N. R., and G. S. Bilotta. 2015. Systematic reviews: separating fact from fiction. Environment International (in press).

Haddaway, N. R., and A. S. Pullin. 2013. Evidence-based conservation and evidence-informed policy: a response to Adams & Sandbrook. Oryx 47:336–338.

Higgins, J. P. T., and S. Green. 2011. Cochrane handbook for systematic reviews of interventions, Version 5.1.0. [updated March 2011]. The Cochrane Collaboration.

Higgs, J., and M. Jones. 2000. Will evidence-based practice take the reasoning out of practice? Pages 307–315 *in* J. Higgs, and M. Jones, editors. Clinical reasoning in the health professionals. Two edition. Butterworth Heineman, Oxford, UK.

Holland, P. W. 1986. Statistics and causal inference. Journal of the American Statistical Association 81:945–960.

Hopkins, J., N. Ockendon, and W. J. Sutherland. 2015. Our mission to transform conservation practice. Conservation Evidence 12:1. http://www.conservationevidence.com/individual-study/5495

Howick, J. 2011. The philosophy of evidence-based medicine. Wiley-Blackwell, Oxford, UK.

Howick, J., P. Glasziou, and J. K. Aronson. 2010. Evidence-based mechanistic reasoning. Journal of the Royal Society of Medicine 103:433–441.

Illari, P. M., F. Russo, and J. Williamson. 2011. Causality in the sciences. Oxford University Press, Oxford, UK.

Karanicolas, P. J., R. Kunz, and G. H. Guyatt. 2008. Evidence-based medicine has a sound scientific base. Chest 133:1067.

Kareiva, P., and M. Marvier. 2012. What is conservation science? BioScience 62:962–969.

Kenward, R. E., et al. 2011. Identifying governance strategies that effectively support ecosystem services, resource sustainability, and biodiversity. Proceedings of the National Academy of Sciences 108:5308–5312.

Kirchner, J. W. 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research 42:1–5.

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

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Lohr, K. N. 2004. Rating the strength of scientific evidence: relevance for quality improvement programs. International Journal for Quality in Health Care 16:9–18.

Mant, R. C., D. L. Jones, B. Reynolds, S. J. Ormerod, and A. S. Pullin. 2013. A systematic review of the effectiveness of liming to mitigate impacts of river acidification on fish and macro-invertebrates. Environmental Pollution 179: 285–293.

Mastrandrea, M., et al. 2010. Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change (IPCC). Available at http://www.ipcc.ch.

Mogas, J., P. Riera, and J. Bennett. 2006. A comparison of contingent valuation and choice modelling with second-order interactions. Journal of Forest Economics 12:5–30.

Morgan, M. G. 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National Academy of Sciences 111:7176–7184.

Mullen, E. J., and D. L. Streiner. 2004. The evidence for and against evidence-based practice. Brief Treatment and Crisis Intervention 4:111–121.

OCEBM Levels of Evidence Working Group. 2011. The Oxford Levels of Evidence 1. URL http://www.cebm.net/index.aspx?o=5653.

Petrokofsky, G., P. Holmgren, and N. D. Brown. 2011. Reliable forest carbon monitoring-systematic reviews as a tool for validating the knowledge base. International Forestry Review 13:56–66.

Petticrew, M., and H. Roberts. 2003. Evidence, hierarchies, and typologies: horses for courses. Theory and Methods 57:527–529.

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

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

Pullin, A. S., and T. M. Knight. 2005. Assessing conservation management's evidence base: a survey of management-plan compilers in the United Kingdom and Australia. Conservation Biology 19:1989–1996.

Rychetnik, L., M. Frommer, P. Hawe, and A. Shiell. 2001. Criteria for evaluation evidence on public health interventions. Journal of Epidemiology and Community Health 56:119–127.

Sackett, D. L., W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. 1996. Evidence based medicine: what it is and what it isn't. Clinical Orthopaedics and Related Research 455:3–5.

Smith, E. P., I. Lipkovich, and K. Ye. 2002. Weight-of-Evidence (WOE): quantitative estimation of probability of impairment for individual and multiple lines of evidence. Human and Ecological Risk Assessment: An International Journal 8:1585–1596.

Smith, R. K., L. V. Dicks, R. Mitchell, and W. J. Sutherland. 2014. Comparative effectiveness research: the missing link in conservation. Conservation Evidence 11:2–6.

Smith, S. D. P., et al. 2015. Rating impacts in a multi-stressor world: a quantitative assessment of 50 stressors affecting the Great Lakes. Ecological Applications 25:717–728.

Spiegelhalter, D. J., and H. Riesch. 2011. Don't know, can't know: embracing deeper uncertainties when analysing risks. Philosophical Transactions of the Royal Society A 369:4730–4750.

Stegenga, J. 2014. Down with the hierarchies. Topoi 33:313–322.

Stewart, G. B., and C. H. Schmid. 2015. Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies. Research Synthesis Methods 6:109–110.

Straus, S. E., P. Glasziou, W. S. Richardson, and R. B. Haynes. 2010. Evidence-based medicine: how to practice and teach it, 4e (Straus evidence-based medicine). Churchill Livingstone.

Sutherland, W. J. 2000. The conservation handbook: research, management and policy. Blackwell Science Ltd.

Sutherland, W. J., and M. A. Burgman. 2015. Use experts wisely. Nature 526:317.

Sutherland, W. J., A. S. Pullin, P. M. Dolman, and T. M. Knight. 2004. Response to Griffiths. Mismatches between conservation science and practice. Trends in Ecology & Evolution 19:565–566.

Sutherland, W. J., A. S. Pullin, P. M. Dolman, and T. M. Knight. 2005. Response to Mathevet and Mauchamp: evidence-based conservation: dealing with social issues. Trends in Ecology & Evolution 20:424–425.

Sutherland, W. J., R. Mitchell, and S. V. Prior. 2012. The role of 'Conservation Evidence' in improving conservation management. Conservation Evidence 9:1–2.

Sutherland, W. J., T. A. Gardner, L. J. Haider, and L. V. Dicks. 2013. How can local and traditional knowledge be effectively incorporated into international assessments? Oryx 48:1–2.

Sutherland, W. J., L. V. Dicks, and R. K. Smith. 2015. What works in conservation? Lessons from conservation evidence. OpenBooks, Cambridge, UK.

Tetlock, P. E. 2005. Expert political judgment: how good is it? How can we know? Princeton University Press, Princeton, New Jersey, USA.

Vetter, D., G. Riicker, and I. Storch. 2013. Meta-analysis: a need for well-defined usage in ecology and conservation biology. Ecosphere 4(6):74.

Walsh, J. C., K. A. Wilson, J. Benshemesh, and H. P. Possingham. 2012. Integrating research, monitoring and management into an adaptive management framework to achieve effective conservation outcomes. Animal Conservation 15:334–336.

Walsh, J. C., L. V. Dicks, and W. J. Sutherland. 2015. The effect of scientific evidence on conservation practitioners management decisions. Conservation Biology 29:88–98.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1890/15-0595/suppinfo