ENVIRONMENTAL EVIDENCE

METHODOLOGY Open Access

Quality assessment tools for evidence from environmental science

Gary S Bilotta[1,2*], Alice M Milner[2,3] and Ian L Boyd[2,4]

Abstract

Assessment of the quality of studies is a critical component of evidence syntheses such as systematic reviews (SRs) that are used to inform policy decisions. To reduce the potential for reviewer bias, and to ensure that the findings of SRs are transparent and reproducible, organisations such as the *Cochrane Collaboration*, the *Campbell Collaboration*, and the *Collaboration for Environmental Evidence*, recommend the use of formal quality assessment tools as opposed to informal expert judgment. However, there is a bewildering array of around 300 formal quality assessment tools that have been identified in the literature, and it has been demonstrated that the use of different tools for the assessment of the same studies can result in different estimates of quality, which can potentially reverse the conclusions of a SR. It is therefore important to consider carefully, the choice of quality assessment tool. We argue that quality assessment tools should: (1) have proven construct validity (i.e. the assessment criteria have demonstrable link with what they purport to measure), (2) facilitate inter-reviewer agreement, (3) be applicable across study designs, and (4) be quick and easy to use. Our aim was to examine current best practice for quality assessment in healthcare and investigate the extent to which these best practices could be useful for assessing the quality of environmental science studies. The feasibility of this transfer is demonstrated in a number of existing SRs on environmental topics. We propose that environmental practitioners should revise, test and adopt the best practice quality assessment tools used in healthcare as a recommended approach for application to environmental science. We provide pilot versions of quality assessment tools, modified from the best practice tools used in healthcare, for application on studies from environmental science.

Keywords: Systematic review, Quality scale, Quality checklist, Bias, Internal validity, External validity, Evidence synthesis

Background

There are significant concerns about the low reproducibility of scientific studies [1,2], and the limited effectiveness of peer-review as a quality-control mechanism [3,4]. Assessment of the quality of scientific studies is critical if they are to be used to inform policy decisions. Quality is a complex concept and the term has been used in different ways; a project using the Delphi consensus method with experts in the field of quality assessment of randomised controlled trials (RCTs) was unable to generate a single definition of quality acceptable to all participants [5]. Nevertheless, from the point of view of a policy-maker, the focus of assessing the quality of a study from environmental science is to establish (i) how near the 'truth' its findings are likely to be, and (ii) how relevant and transferable the findings are to the particular setting or population of policy interest. It is important to assess these aspects of study quality separately, as each component has different implications on how the findings of a study should be interpreted.

The first of these aspects of concern to policy-makers is referred to as the internal validity of the study; it is a measure of the extent to which a study is likely to be free from bias. A bias is a systematic error resulting from poor study design or issues associated with conduct in the collection, analysis, interpretation, and reporting of data [6]. Biases can operate in either direction, which if unaccounted for may ultimately affect the validity of the conclusions from a study [7]. Biases can vary in magnitude: some are small and trivial, and some are substantial such that an apparent finding may be entirely due to bias [7]. It is usually impossible to know the extent to which biases have affected the results of a particular study, although there is good empirical evidence that

* Correspondence: g.s.bilotta@brighton.ac.uk
[1]University of Brighton, Brighton BN2 4GJ, UK
[2]Department for Environment, Food and Rural Affairs, London SW1P 3JR, UK
Full list of author information is available at the end of the article

particular flaws in the design, conduct and analysis of studies lead to an increased likelihood of bias [8]. This empirical evidence can be used to support assessments of internal validity.

The second of these aspects of concern to policy-makers is referred to as the external validity of the study. The assessment of a study's external validity, depends partly on the purpose for which the study is to be used, and is less relevant without internal validity [7]. External validity is closely connected with the generalisability or applicability of a study's findings. Its assessment often considers the directness of the evidence (i.e. the level of similarity between the population/environment/ecosystem studied and the population/environment/ecosystem of policy interest, and the level of similarity between the intervention/treatment conditions and the temporal and spatial scales of the study in relation to the situation of policy interest), and the precision of the study (in this case imprecision refers to *random error*, meaning that multiple replications of the same study will produce different findings because of sampling variation) [7]. Another component of external validity is the statistical conclusion validity, i.e. the degree to which conclusions about the relationship among variables based on the data are correct or 'reasonable' [9]. Statistical conclusion validity concerns the features of the study that control for Type I errors (finding a difference or correlation when none exists) and Type II errors (finding no difference when one exists). Such controls include the use of adequate sampling procedures, appropriate statistical tests, and reliable measurement procedures.

Assessments of study quality can be made informally using expert judgement. However, experts in a particular area frequently have pre-formed opinions that can bias their assessments [10-12]. In order to reduce the potential for reviewer bias and to ensure that the findings of a SR are transparent and reproducible, organisations such as the *Cochrane Collaboration* (who prepare, maintain and disseminate SRs on the effectiveness of healthcare interventions), the *Campbell Collaboration* (who prepare, maintain, and disseminate SRs on the effectiveness of social and behavioural interventions in education, social welfare, and crime and justice), and the *Collaboration for Environmental Evidence* (CEE - who prepare, maintain and disseminate SRs on environmental interventions), recommend the use of formal quality assessment tools, recognising that the merits of a formal approach outweigh the drawbacks.

Around 300 formal study quality assessment tools have been identified in the literature [13]. These tools are designed to provide a means of objectively assessing the overall quality of a study using itemised criteria, either qualitatively in the case of checklists or quantitatively in the case of scales [14]. However, perhaps unsurprisingly given the diverse range of criteria included within quality assessment tools, it has been empirically demonstrated that the use of different quality tools for the assessment of the same studies results in different estimates of quality, which can potentially reverse the conclusions of a SR and therefore potentially lead to misinformed policies [15-17]. For example, in the healthcare field, a meta-analysis of 17 trials comparing the effectiveness of low-molecular-weight heparin (LMWH) with standard heparin for prevention of post-operative thrombosis, trials that were identified as 'high quality' by some of the 25 quality scales interrogated, indicated that LWMH was not superior to standard heparin, whereas trials identified as 'high quality' by other scales led to the opposite conclusions - that LMWH was beneficial [18].

It is therefore very important to consider carefully, the choice of quality assessment tool to be used in SRs [19]. At present, the CEE (2013) does not specify which quality assessment tool to use in SRs on environmental topics. Authors of CEE SRs are permitted to use any quality assessment tool as a basis for their specific exercise, but they should either explain why they used them as such (no modification, because not considered to be needed, and why), or adapt them to their own review, in which case the decisions made must be stated and justified [20]. This stance on the use of quality assessment tools may leave readers and users of their reviews to question the reliability of review findings. We argue that in order to satisfy their purpose, quality assessment tools should possess four features described in the Desirable features of a quality assessment tool section.

Desirable features of a quality assessment tool

(I) The tool should have construct validity (i.e. the included criteria measure what they purport to be measuring)

Some of the existing quality assessment tools have been criticised for (i) the lack of rationale for the criteria used, (ii) inclusion of criteria that are unlikely, or not proven, to be related to internal or external validity, and (iii) unjustified weighting for each of the criteria used [21,22]. Inclusion of non-relevant criteria in assessment tools can dilute or distort the relevant criteria resulting in assessments that have minimal correlation with the aspects of study quality that matter to policy-makers (i.e. internal and external validity). Empirical evidence about the importance of study design features in reducing the risk of bias has accumulated rapidly since the late 1990s [8]. This evidence has mainly been obtained by a powerful but simple technique of investigating variations in the results of studies of the same intervention according to features of their study design [23]. The process involves first identifying substantial numbers of studies both with and without the design feature of interest. Results are then compared between the studies fulfilling and not fulfilling

each design criterion, to obtain an estimate of the systematic bias removed by the design feature.

Assessment of the external validity of studies is more likely to require situation-specific criteria; for example the spatial and temporal scale of studies may be particularly important aspects to consider to determine the generalisability of environmental studies. Nevertheless, criteria should only be included if there is strong empirical evidence to support their implementation.

(II) The tool should facilitate good inter-reviewer agreement

The purpose of using a formal quality assessment tool, as opposed to informal expert judgement, is to reduce the potential for reviewer bias and to ensure that the assessment is transparent and reproducible. It is therefore essential that the tool used facilitates inter-reviewer agreement (i.e. that assessments are reproducible using different reviewers); otherwise there is little advantage over an informal expert judgement system. Inter-reviewer agreement should be tested across a range of studies during tool development, but also should be checked during the conduct of each SR to ensure that the reviewers are interpreting the tool correctly with regards to their specific review topic. Surprisingly, the inter-reviewer reliability of tools is not always assessed when tools are developed. For example, of the 60 'top tools' identified by Deeks et al. [19], only 24 tools were tested for their inter-reviewer reliability during development. When inter-reviewer agreement is tested, common statistical measures include Kappa statistics (which adjust the proportion of records for which there was agreement, by the amount of agreement expected by chance alone), and/or correlation coefficients. The level of inter-reviewer agreement is influenced by the design of the quality assessment tool (clarity, relevance to the study being assessed, and degree of subjectivity), but is also dependent on the experience of the reviewers; de Oliveira et al. [24] achieved higher agreement for experienced reviewers, whereas Coleridge Smith [25] found higher agreement for two methodologists combined rather than clinicians alone or for a clinical/methodologist pairing.

(III) The tool should be applicable across study designs

If reviewers can only assess one study design type (e.g. RCTs) with a given tool, multiple tools will be required for any SR that includes multiple study design types. The use of multiple tools increases the complexity of the review process and complicates the interpretation of the findings of the SR for the readership, including policy-makers. Furthermore, if authors are free to pick and choose which tool to use for a given study design, this opens up the process to reviewer bias whereby the reviewer may select a tool which emphasises research or studies that meet their own opinions, prejudices or commercial interests. A solution to this is to use a tool that is capable of assessing quality across different study design types. Downs and Black [26] argue that although different study design types have fundamental differences, common factors are measured: the intervention, potential confounders, and the outcome. Many study designs test for an association between the intervention and the outcome and aim to minimise flaws in the design that will bias the measurement of an association. The vulnerability of each design to different biases varies but the kind of biases that the study designs seek to exclude are the same, and therefore it is possible for a quality assessment tool to be applicable across study designs. It is important that any 'universal' quality assessment tools have clear guidance to aid interpretation of the criteria and to prevent erroneous classification when applied to different study designs. Furthermore, it is useful if aspects of the tool can be 'switched off' if not relevant to a particular study design being evaluated. Generic tools can be frustrating to use if they do not apply to the case in hand and they may suffer from lower inter-reviewer reliability if significant subjective interpretation is required when applying generic criteria to specific cases.

(IV) The tool should be quick and easy to use

It is common for SRs to cite a significant number of studies. It is therefore important from a practical point of view, for any quality assessment tool that is designed for use in SRs, to be quick and easy to use. The time taken to make an assessment with a given tool depends on the number of criteria that the tool is comprised of, and the degree of decision-making required to answer each criterion (dichotomous response or a more complex written response). The number of criteria in the 193 quality assessment tools reviewed by Deeks et al. [19], ranged from 3 to 162, with a median of 12. The tools selected as the 'top 60 tools' or the 'best 14 tools' according to the benchmarks laid out by Deeks et al. [19], had a higher median number of criteria compared with the unselected tools, and took an average time to complete of between 5 to 30 minutes per assessment.

Current best practice in healthcare

The Cochrane Collaboration, who are internationally renowned for their SRs in healthcare, have recently developed an approach to quality assessment that satisfies these four criteria. Until 2008, the Cochrane Collaboration used a variety of quality assessment tools, mainly checklists in their SRs [27]. However, owing to knowledge of inconsistent assessments using different tools to assess the same studies [18], and to growing criticisms of many of these tools [28], in 2005 the Cochrane Collaboration's Methods Groups (including statisticians,

epidemiologists, and review authors) embarked on developing a new evidence-based strategy for assessing the quality of studies, focussing on internal validity [7]. The resultant product of this research was the Cochrane Collaboration's Risk of Bias Tool (Risk of Bias Tool) [7]. The Risk of Bias Tool is used to make an assessment of internal validity, or risk of bias, of a study through consideration of the following five key aspects (domains) of study design that are empirically proven to control risk of bias. (1) **randomization** (minimises the risk of *selection bias*[a]) [29-31], (2) **allocation concealment and** (3) **blinding** (minimises the risk of *performance bias*[b] and *detection bias*[c] due to participants' or investigators' expectations) [15,22,32-37], (4) **follow-up** (high follow-up of participants from enrolment to study completion minimises the risk of *attrition bias*[d]) [33,34], and (5) **unselective reporting** of outcomes (minimises the risk of *reporting bias*[e]) [38,39]. The Risk of Bias Tool provides criteria, shown in Additional file 1, to guide the assessment of each of these domains, classifying each domain as low, high or unclear risk of bias.

Reviewers must provide support for judgment in the form of a succinct free text description or summary of the relevant trial characteristic on which assessments of risk of bias are based. This is designed to ensure transparency in how assessments are reached. The Cochrane Collaboration recommends that for each SR, judgments must be made independently by at least two people, with any discrepancies resolved by discussion in the first instance [7]. Some of the items in the tool, such as methods for randomisation, require only a single assessment for each trial included in the review. For other items, such as blinding and incomplete outcome data, two or more assessments may be used because they generally need to be made separately for different outcomes (or for the same outcome at different time points) [7]. The classification for each domain is presented for all studies in a manner shown in Additional file 2. To draw conclusions about the overall risk of bias for an outcome it is necessary to summarise these domains. Any assessment of the overall risk of bias involves consideration of the relative importance of different domains. Review authors will have to make assessments about which domains are most important in the current review [7]. For example, for highly subjective outcomes such as pain, authors may decide that blinding of participants (i.e. where information about the test that might lead to bias in the results is concealed from the participants) is critical. How such assessments are reached should be made explicit and they should be informed by consideration of the empirical evidence relating each domain to bias, the likely direction of bias, and the likely magnitude of bias. The Cochrane Collaboration Handbook provides guidance to support summary assessments of the risk of bias, but a suggested simple approach is that a low risk of bias classification is given to studies which have a low risk of bias for all key domains; a high risk of bias classification is given to studies which have a high risk of bias for one or more key domains; and an unclear risk of bias classification is given to studies which have an unclear risk of bias for one or more key domains.

The Risk of Bias Tool was published in February 2008 and was adopted as the recommended method throughout the Cochrane Collaboration. A three stage project to evaluate the tool was initiated in early 2009 [7]. This evaluation project found that the 2008 version of the Risk of Bias Tool took longer to complete than previous methods. Of 187 authors surveyed, 88% took longer than 10 minutes to complete the new tool, 44% longer than 20 minutes, and 7% longer than an hour, but 83% considered the time taken acceptable [7]. An independent study (cross sectional analytical study on a sample of 163 full manuscripts of RCTs in child health), also found that the 2008 version of the Risk of Bias Tool took longer to complete than some other quality assessment tools (the investigators took a mean of 8.8 minutes per person for a single predetermined outcome using the Risk of Bias tool compared with 1.5 minutes for a previous rating scale for quality of reporting) [40]. The same study reported that inter-reviewer agreement on individual domains of the Risk of Bias Tool ranged from slight ($\kappa = 0.13$) to substantial ($\kappa = 0.74$), and that agreement was generally poorer for those items that required more judgment. An interesting finding from the Hartling et al. [40] study was that the overall risk of bias as assessed by the Risk of Bias Tool differentiated effect estimates, with more conservative estimates for studies assessed to be at low risk of bias compared to those at high or unclear risk of bias. On the basis of the evaluation project, the first (2008) version of the Risk of Bias Tool was modified to produce a second (2011) version, which is the version shown in Additional file 1.

The Risk of Bias Tool enables reviewers to assess the internal validity of primary studies. Internal validity is only one aspect of study quality that decision makers should be interested in. In order to assess the overall quality of a body of evidence, the Cochrane Collaboration use a system developed by the Grades of Recommendation, Assessment, Development and Evaluation (GRADE) Working Group [41-44]. This approach is now used by the World Health Organisation (WHO) and the UK National Institute for Health and Care Excellence (NICE) amongst 20 other bodies internationally [7]. For purposes of SRs, the GRADE approach describes four levels of quality. The highest quality rating is initially for RCT evidence (see Table 1). Review authors can, however, downgrade evidence to moderate, low, or even very low quality evidence, depending on the presence of the five

Table 1 Classification of study designs

Study design	Description
Randomised Controlled Trial (RCT)	RCTs are studies in which study units are randomly allocated to intervention or control groups and followed up over time to assess any differences in outcome rates. Randomisation with allocation concealment (double blinded) ensures that on average known and unknown determinants of outcome (including confounding factors) are evenly distributed between groups.
Observational Study	Observational studies are studies in which natural variation in interventions (or exposure) among study units is investigated to explore the effect of the interventions (or exposure) on outcomes. These include designs such as:
	• **Controlled before-and-after study**: A follow-up study of study units that have received an intervention and those that have not, measuring the outcome variable both at baseline and after the intervention period, comparing either final values if the groups are comparable at baseline, or change in scores if they were not.
	• **Concurrent cohort study**: A follow-up study that compares outcomes between study units that have received an intervention and those that have not. Study units are studied during the same (concurrent) period either prospectively or, more commonly, retrospectively.
	• **Historical cohort study**: A variation on the traditional cohort study where the outcome from a new intervention is established for study units studied in one period and compared with those that did not receive the intervention in a previous period, i.e. study units are not studied concurrently.
	• **Case–control study**: Study units with and without a given outcome are identified (cases and controls respectively) and exposure to a given intervention(s) between the two groups compared.
	• **Before-and-after study**: Comparison of outcomes from study units before and after an intervention is introduced. The before and after measurements may be made in the same study units, or in different samples.
	• **Cross-sectional study**: Examination of the relationship between outcome and other variables of interest as they exist in a defined population at one particular time point.
Case Study	Case studies are studies which describe a number of cases of an intervention and outcome, with no comparison against a control group.

factors described in detail in Schünemann et al. [45]: (1) risk of bias across studies, assessed using the Cochrane Collaboration's Risk of Bias Tool, (2) directness of evidence (i.e. the extent to which the study investigates a similar population, intervention, control, outcome), (3) heterogeneity in the findings, (4) precision of effect estimates, and (5) risk of publication bias (i.e. systematic differences in the findings of published and unpublished studies). Usually, quality rating will fall by one level for each factor, up to a maximum of three levels for all factors. If there are very severe problems for any one factor (e.g. when assessing limitations in design and implementation, all studies were unconcealed, unblinded, and lost over 50% of their participants to follow-up), evidence may fall by two levels due to that factor alone. Review authors will generally grade evidence from sound observational studies (see Table 1) as low quality. These studies can be upgraded to moderate or high quality if: (1) such studies yield large effects and there is no obvious bias explaining those effects, (2) all plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results show no effect, and/or (3) if there is evidence of a dose–response gradient. Guidance and justification for making these assessments is provided by Schünemann et al. [45]. The very low quality level includes, but is not limited to, studies with critical problems and unsystematic observations (e.g. case series/reports - see Table 1).

Additional file 3 presents the decisions that must be made in going from assessments of the risk of bias (using the Risk of Bias Tool) to assessments about study limitations for each outcome included in a 'Summary of findings' table. As can be seen from this table, a rating of high quality evidence can be achieved only when most of the evidence comes from studies that met the criteria for low risk of bias (as determined using the Risk of Bias Tool).

The developers of the GRADE approach caution against a completely mechanistic approach toward the application of the criteria for rating the overall quality of the evidence up or down; arguing that although GRADE suggest the initial separate consideration of five categories of reasons for rating down the quality of evidence, and three categories for rating up, with a yes/no decision regarding rating up or down in each case, the final rating of overall evidence quality occurs in a continuum of confidence in the validity, precision, consistency, and applicability of the estimates. Fundamentally, the assessment of evidence quality is a subjective process and GRADE should not be seen as obviating the need for or minimising the importance of judgment or as suggesting that quality can be objectively determined. The use of GRADE will not guarantee consistency in assessment. There will be cases in which competent reviewers will have honest and legitimate disagreement about the interpretation of evidence. In such cases, the merit of GRADE is that it provides a framework that guides reviewers through the critical components of the assessment and an approach to analysis and communication that encourages transparency and an explicit accounting of the assessments. Testing of inter-reviewer agreement

between blinded reviewers using a pilot version of GRADE showed that there was a varied amount of agreement on the quality of evidence for individual outcomes (kappa coefficients for agreement beyond chance ranged from 0 to 0.82) [46,47]. Nevertheless, most of the disagreements were easily resolved through discussion (GRADE assessments are conducted by at least two reviewers who are blinded before agreeing assessments). In general, reviewers of the pilot version found the GRADE approach to be clear, understandable and sensible [46]. On the basis of the evaluation research, modifications were made to the GRADE approach to improve inter-reviewer agreement [46,47].

Applicability to environmental studies

The Cochrane Collaboration's development of a single recommended tool to assess internal validity of primary studies (Risk of Bias Tool), combined with the use of the GRADE system to rate the overall quality of a body of evidence (considering factors affecting external validity), has promoted reproducible and transparent assessments of quality across the Cochrane Collaboration's SRs. The question is, can the same be done for quality assessment of environmental studies?

Without significant trials it is difficult to know the extent to which tools, such as the Risk of Bias Tool and the GRADE Tool, could be adopted and applied, usefully, across different types of environmental studies. There are obvious similarities between healthcare studies and studies on animal and plant health that form part of the environmental evidence-base, but for other topics in environmental science, the similarities are perhaps less immediately obvious. Nevertheless, there are precedents set that demonstrate the potential transferability of tools from healthcare to a range of environmental studies. For example:

There are six environmental SRs [48-53] that have used a hierarchy of evidence approach which was developed by Pullin and Knight [54], based on an adaptation of early hierarchical approaches used in healthcare. This approach is not too distant from the GRADE approach, although the Pullin and Knight [54] evidence hierarchy includes expert opinion as a form of evidence (GRADE does not do this), and in terms of application, the assessments with the Pullin and Knight [54] hierarchy are made on individual studies, not the overall body of evidence for each outcome (as is the case with GRADE). The use of an evidence hierarchy alone is not well justified: using this approach assumes that a study design at a higher level in the hierarchy is methodologically superior to one at a lower level and that studies at the same level are equivalent, but this ranking system is far too simplistic given that there are many design characteristics that can comprise a given study [55]. It is well known that an RCT (which under this system would be rated as the highest quality form of evidence) can suffer from bias if it has poor randomisation, no allocation concealment, no blinding of participants or investigators, uneven reporting of outcomes, or high unexplained attrition. These features can make an RCT more prone to bias than a well-designed observational study. This issue is explicitly recognised in the combined GRADE and Risk of Bias Tool approach used by the Cochrane Collaboration, but not explicitly recognised in the Pullin and Knight [54] approach described above.

There are a further ten environmental SRs [56-65], that have used an approach to quality assessment that is similar to the current best practice in healthcare, albeit using the Pullin and Knight [54] hierarchy of evidence in combination with an assessment of some of the individual aspects of study design that are empirically proven to be controls on selection bias, performance bias, detection bias and attrition bias. Although these environmental versions of the Cochrane Collaboration's current approach are: untested outside of each of the review teams, designed to be review-specific, use an evidence hierarchy with the limitations mentioned above, and make an assessment of overall quality for each study (rather than for the body of evidence overall for each outcome) - the general approach is very similar, demonstrating that it is feasible to capitalise on the research and development that the healthcare field has completed on quality assessment.

Nevertheless, some changes to the healthcare tools are necessary. Table 2 (Environmental-Risk of Bias Tool) provides an example of the criteria that could be used to guide the assessment of the internal validity of environmental studies. The original clinical terminology from the Cochrane Collaboration's Risk of Bias Tool has been removed and replaced with scientific terminology where appropriate. We have also added definitions of all of the sources of bias, as subtitles in Table 2, to facilitate understanding by the environmental community.

Table 3 (Environmental-GRADE Tool) provides an example of the criteria that could be used to guide the assessment of the external validity of environmental studies. We modified this from the original GRADE approach so that the tool now provides an assessment of the quality of individual studies rather than the overall body of evidence. This modification was deemed necessary owing to the apparent higher diversity of study designs used in environmental science, which could make it difficult to acquire a meaningful summary of quality across studies. On the basis of this change in the GRADE assessment, we argue that the consideration of (i) heterogeneity of study findings and (ii) publication bias, MUST be conducted separately during the SR process. Reviewers are advised to consider the use of forest plots and funnel plots to assess (i) heterogeneity of

Table 2 Criteria for the assessment of internal validity of a study, using the Environmental-Risk of Bias Tool, adapted from the Cochrane Collaboration's Risk of Bias Tool

SELECTION BIAS DUE TO INADEQUATE RANDOMISATION	
Selection bias refers to systematic differences between baseline characteristics of the study units that are to be compared [66]. The likelihood of selection bias can be minimised through randomisation of treatments to study units and through allocation sequence concealment [66]. Proper randomisation requires that a rule for allocating treatments to study units must be specified based on some chance (random) process [66]. This is referred to as the sequence generation [66].	
Criteria for an assessment of 'Low risk' of bias.	The investigators describe a random component in the sequence generation process such as: • Referring to a random number table; • Using a computer random number generator; • Coin tossing; • Shuffling cards or envelopes; • Throwing dice; • Drawing of lots; • Minimisation[1]
Criteria for an assessment of 'High risk' of bias.	The investigators describe a non-random component in the sequence generation process. Usually, the description would involve some systematic, non-random approach, for example: • Sequence generated by odd or even date of birth of study units, or odd or even latitude of sites; • Sequence generated by some rule based on site number or site code in a database; Other non-random approaches involving judgement or some method of non-random allocation of study units, for example: • Allocation by judgement of the investigator; • Allocation based on the results of a laboratory test or a series of tests; • Allocation by availability of the intervention.
Criteria for an assessment of 'Unclear risk' of bias.	Insufficient information about the sequence generation process to permit assessment of 'Low risk' or 'High risk'.
SELECTION BIAS DUE TO INADEQUATE ALLOCATION CONCEALMENT	
Selection bias refers to systematic differences between baseline characteristics of the study units that are to be compared [66]. The likelihood of selection bias can be minimised through randomisation of treatments to study units and through allocation sequence concealment [66]. Proper allocation concealment involves taking steps to secure strict implementation of that schedule of random assignments by preventing foreknowledge (by study units and investigators) of the forthcoming allocations [66]. This process if often termed allocation concealment, although it could more accurately be described as allocation sequence concealment [66].	
Criteria for an assessment of 'Low risk' of bias.	Study units and investigators enrolling study units could not foresee allocations because one of the following, or an equivalent method, was used to conceal allocation: • Central allocation (including telephone, web-based and database controlled randomisation); • Sequentially numbered treatment containers of identical appearance (e.g. test of effectiveness of different pesticides, where different pesticides appear identical);
Criteria for an assessment of 'High risk' of bias.	Study units or investigators enrolling study units could possibly foresee allocations and thus introduce selection bias, such as allocation based on: • Using an open random allocation schedule (e.g. a list of random numbers); • Alternation or rotation; • Date of birth of participants; • Case record number/site code; • Any other explicitly unconcealed procedure.
Criteria for an assessment of 'Unclear risk' of bias.	Insufficient information to permit assessment of 'Low risk' or 'High risk'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite assessment.
PERFORMANCE BIAS	
Performance bias refers to systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest [66]. After enrolment into the study, blinding (or masking) of study units and personnel (assuming these are different to the investigators responsible for allocation of study units to study groups) may reduce the risk that knowledge of which intervention was received, rather than the intervention itself, affects outcomes [66]. Effective blinding can also ensure that the compared groups receive a similar amount of attention, ancillary treatment and diagnostic investigations [66].	

Table 2 Criteria for the assessment of internal validity of a study, using the Environmental-Risk of Bias Tool, adapted from the Cochrane Collaboration's Risk of Bias Tool *(Continued)*

Criteria for an assessment of 'Low risk' of bias.	Any one of the following: • No blinding or incomplete blinding, but the review authors judge that the outcome is not likely to be influenced by lack of blinding; • Blinding of study units and key study personnel ensured, and unlikely that the blinding could have been broken.
Criteria for an assessment of 'High risk' of bias.	Any one of the following: • No blinding or incomplete blinding, and the outcome is likely to be influenced by lack of blinding; • Blinding of key study units and personnel attempted, but likely that the blinding could have been broken, and the outcome is likely to be influenced by lack of blinding.
Criteria for an assessment of 'Unclear risk' of bias..	Any one of the following: • Insufficient information to permit assessment of 'Low risk' or 'High risk'; The study did not address this outcome

DETECTION BIAS

Detection bias refers to systematic differences between groups in how outcomes are determined [66]**.** Blinding of outcome assessors may reduce the risk that knowledge of which intervention was received, rather than the intervention itself, affects outcome measurement [66]. Blinding of outcome assessors can be especially important for assessment of subjective outcomes [66].

Criteria for an assessment of 'Low risk' of bias.	Any one of the following: • No blinding of outcome assessment, but the review authors judge that the outcome measurement is not likely to be influenced by lack of blinding; • Blinding of outcome assessment ensured, and unlikely that the blinding could have been broken.
Criteria for an assessment of 'High risk' of bias.	Any one of the following: • No blinding of outcome assessment, and the outcome measurement is likely to be influenced by lack of blinding; • Blinding of outcome assessment, but likely that the blinding could have been broken, and the outcome measurement is likely to be influenced by lack of blinding.
Criteria for an assessment of 'Unclear risk' of bias.	Any one of the following: • Insufficient information to permit assessment of 'Low risk' or 'High risk'; The study did not address this outcome.

ATTRITION BIAS DUE TO INCOMPLETE OUTCOME DATA

Attrition bias refers to systematic differences between groups in withdrawals from a study [66]. Withdrawals from the study lead to incomplete outcome data [66]. There are two reasons for withdrawals or incomplete outcome data in clinical trials [66]. Exclusions refer to situations in which some study units are omitted from reports of analyses, despite outcome data being available to the trialists [66]. Attrition refers to situations in which outcome data are not available [66].

Criteria for an assessment of 'Low risk' of bias.	Any one of the following: • No missing outcome data; • Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to be introducing bias); • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have a relevant impact on the intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have a relevant impact on observed effect size; • Missing data have been imputed using appropriate methods.
Criteria for an assessment of 'High risk' of bias.	Any one of the following: • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce relevant bias in intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce relevant bias in observed effect size; • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization; • Potentially inappropriate application of simple imputation.

Table 2 Criteria for the assessment of internal validity of a study, using the Environmental-Risk of Bias Tool, adapted from the Cochrane Collaboration's Risk of Bias Tool *(Continued)*

Criteria for an assessment of 'Unclear risk' of bias.	Any one of the following: • Insufficient reporting of attrition/exclusions to permit assessment of 'Low risk' or 'High risk' (e.g. number randomized not stated, no reasons for missing data provided); The study did not address this outcome.
REPORTING BIAS DUE TO SELECTIVE REPORTING	
Reporting bias refers to systematic differences between reported and unreported findings [66]. Within a published report those analyses with statistically significant differences between treatment groups are more likely to be reported than non-significant differences [66]. This sort of 'within-study publication bias' is usually known as outcome reporting bias or selective reporting bias, and may be one of the most substantial biases affecting results from individual studies [67].	
Criteria for an assessment of 'Low risk' of bias.	Any of the following: • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way; • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon).
Criteria for an assessment of 'High risk' of bias.	Any one of the following: • Not all of the study's pre-specified primary outcomes have been reported; • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified; • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis; • The study report fails to include results for a key outcome that would be expected to have been reported for such a study.
Criteria for an assessment of 'Unclear risk' of bias.	Insufficient information to permit assessment of 'Low risk' or 'High risk'. It is likely that the majority of studies will fall into this category.
OTHER BIAS	
In addition there are other sources of bias that are relevant only in certain circumstances [66]. These relate mainly to particular study designs (e.g. carry-over in cross-over trials and recruitment bias in cluster-randomized trials); some can be found across a broad spectrum of trials, but only for specific circumstances (e.g. contamination, whereby the experimental and control interventions get 'mixed'); and there may be sources of bias that are only found in a particular setting [66].	
Criteria for an assessment of 'Low risk' of bias.	The study appears to be free of other sources of bias.
Criteria for an assessment of 'High risk' of bias.	There is at least one important risk of bias. For example, the study: • Had a potential source of bias related to the specific study design used; or • Has been claimed to have been fraudulent; or • Had some other problem.
Criteria for an assessment of 'Unclear risk' of bias.	There may be a risk of bias, but there is either: • Insufficient information to assess whether an important risk of bias exists; or • Insufficient rationale or evidence that an identified problem will introduce bias.

[1]Minimisation is a method that seeks to limit any difference in the distribution of known or suspected determinants of outcome, so that any effect can be attributed to the treatment under test [68]. The investigators determine at the outset which factors they would like to see equally represented in the study groups [68]. The treatment allocation is then made, not purely by chance, but by determining in which group inclusion of the patient would minimise any differences in these factors [68]. The allocation may rely on minimisation alone, or still involve chance but "with the dice loaded" in favour of the allocation which minimises the differences [68].

study findings and (ii) publication bias, respectively. The Cochrane Collaboration's current Review Manager software (available free-of-charge from: http://tech.cochrane.org/revman/download), can create these plots relatively easily if the appropriate data are available. The assessment of study quality through use of the Environmental-Risk of Bias and Environmental-GRADE tools should aid reviewers in their interpretation of heterogeneity of study findings and publication bias, but not vice-versa.

As illustrated in Table 3, the highest quality rating is initially for RCT evidence. Review authors can, however, downgrade RCT evidence to moderate, low, or even very

Table 3 Criteria for assessing the overall quality of an environmental study, using the Environmental-GRADE Tool, adapted from the GRADE approach used by the Cochrane Collaboration (Balshem et al. [69]; Schünemann et al. [45]; Guyatt et al. [70])

Underlying methodology	Quality rating	Interpretation
RCT; or double-upgraded observational study.	**High**	We are very confident that the true effect estimate lies close to that of the estimate of the effect.
Downgraded RCT; or upgraded observational study.	**Moderate**	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Double-downgraded RCT; or observational study.	**Low**	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.
Triple-downgraded RCT; or downgraded observational study; or case study.	**Very low**	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.

low quality evidence, depending on the presence of the three factors discussed below:

(I) The risk of bias within the study

This is assessed using the Environmental-Risk of Bias Tool, criteria to guide this assessment are provided in Table 2. For making summary assessments of risk of bias for each study, the low risk of bias classification is given to studies that have low risk of bias for all key domains; the high risk of bias classification is given to studies that have a high risk of bias for one or more key domains; the unclear risk of bias is given to studies in which the risk of bias is uncertain for one or more key domains. The principle of considering risk of bias for observational studies is exactly the same as for RCTs. However, potential biases are likely to be greater for observational studies. Review authors need to consider (a) the weaknesses of the designs that have been used (such as noting their potential to ascertain causality), (b) the execution of the studies through a careful assessment of their risk of bias, especially (c) the potential for selection bias and confounding to which all non-randomised studies are susceptible and (d) the potential for reporting biases, including selective reporting of outcomes.

(II) The directness of the study

The directness of the study refers to the extent to which the study: investigates a similar species, population, community, habitat or ecosystem as that of policy interest; investigates a similar application of the intervention/treatment as that of policy interest; investigates the phenomenon at a similar spatial scale as that of policy interest; investigates the phenomenon at a similar temporal scale as that of policy interest. Reviewers should make assessments transparent when they believe downgrading is justified based on directness of evidence. These assessments should be supported with significant empirical evidence.

(III) The precision of effect estimates

In this case imprecision refers to *random error*, meaning that multiple replications of the same study would produce different effect estimates because of sampling variation. When studies have small sample sizes (considering the amount of group variability and the reliability of the outcome measures) the risk of imprecision increases. In these circumstances reviewers can lower their rating of the quality of the evidence. In order for reviewers to make this assessment, they will need information about the uncertainties associated with the study design, including the precision of sampling and measurements. Related to this is the statistical conclusion validity, which as described in Background section of this paper, is the degree to which conclusions about the relationship among variables based on the data are correct or 'reasonable'. Statistical conclusion validity concerns the features of the study that control for Type I and Type II errors. Such controls include the use of adequate sampling procedures, appropriate statistical tests, and reliable measurement procedures. The most common sources of threats to statistical conclusion validity are low statistical power, violated assumptions of the test statistics[f], fishing and the error rate problem[g], unreliability of measures[h], and restriction of range[i]. There are recognised statistical methods to test for common causes of violated assumptions of test statistics (e.g. normal distribution tests). If the study does not report the results of these statistical checks, it would be beneficial to try and obtain the statistical test results from the original authors. Reviewers should make assessments transparent when they believe downgrading is justified based on the precision of effect estimates. These assessments should be supported with significant empirical evidence and guided through consultation with a statistical expert.

Usually, quality rating will fall by one level for each factor, up to a maximum of three levels for all factors. If there are very severe problems for any one factor (e.g. when assessing limitations in design and implementation, all studies were unconcealed, unblinded, and lost over 50% of their study units to follow-up), RCT evidence may fall by two levels due to that factor alone. Reviewers will generally grade evidence from sound observational studies as

low quality. These studies can be upgraded to moderate or high quality if (1) such studies yield large effects and there is no obvious bias explaining those effects, (2) all plausible confounding factors would reduce a demonstrated effect or suggest a spurious effect when results show no effect, and/or (3) if there is evidence of a dose–response gradient. Reviewers should make assessments transparent when they believe upgrading is justified based on these factors. These assessments should be supported with empirical evidence where available to the reviewers; this can be sourced both from evidence presented in the study, or presented in other studies. The very low quality level includes, but is not limited to, studies with critical problems and unsystematic observations (e.g. case studies).

Conclusions

This article examined the current best practices for quality assessment in healthcare (Cochrane Collaboration's Risk of Bias Tool and the GRADE system) and investigated the extent to which these practices/tools could be useful for assessing the quality of environmental science. We highlighted that the feasibility of this transfer has been demonstrated in a number of existing SRs on environmental topics. It is therefore not difficult to imagine that the Cochrane Collaboration's Risk of Bias Tool and the GRADE system could be adapted and applied routinely, as a preferred method, as part of the quality assessment of environmental science studies cited in CEE SRs. We propose a pilot version of the Environmental-Risk of Bias Tool and the Environmental-GRADE Tool for this purpose, and we provide worked examples in Additional files 4, 5 and 6.

Some of the terminology used in the original Risk of Bias Tool has been changed from clinical terms to environmental science terms. Definitions of all of the sources of bias have been added as subtitles to the Risk of Bias assessment criteria. The original GRADE Tool has been modified more substantially so that the Environmental-GRADE Tool now provides an assessment of the quality of individual studies rather than the overall body of evidence. This modification was deemed necessary because of the higher diversity of study designs used in environmental science, which could make it difficult to acquire a meaningful summary of quality across studies. On the basis of this change in assessment, we argue that the consideration of: (i) heterogeneity of study findings, and (ii) publication bias, MUST be conducted during the evidence synthesis and meta-analysis stages of the SR process.

We suggest that once used in a number of environmental SRs, it will be possible to conduct an evaluation of the Environmental-Risk of Bias Tool and Environmental-GRADE Tool to understand how ease of use, applicability across study designs, and degree of inter-reviewer agreement could be enhanced. This 'learning by doing' application of the pilot versions of these quality assessment tools is exactly how the Cochrane Collaboration and the GRADE Working Group refined their respective tools. The Cochrane Collaboration and GRADE Working Group continue to refine their quality assessment tools as more evidence comes forward that supports this. The environmental community should capitalise on this ongoing methodological research and development, harmonising its tools where appropriate. At the same time, there is a need to build capacity in methodological expertise within the environmental field. Finally, the environmental science community and research funding organisations should use the criteria contained within the Environmental-Risk of Bias Tool and Environmental-GRADE Tool to enhance the quality of funded studies in the future.

Endnotes

[a]Systematic differences between baseline characteristics of the groups that are to be compared.

[b]Systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest due to lack of blinding of investigators.

[c]Systematic differences between groups in how outcomes are determined.

[d]Systematic differences between groups in withdrawals from a study/loss of samples.

[e]Systematic differences between reported and unreported findings.

[f]Most statistical tests involve assumptions about the data that make the analysis suitable for testing a hypothesis. Violating the assumptions of statistical tests can lead to incorrect inferences about the cause-effect relationship. Violations of assumptions may make tests more or less likely to make Type I or II errors.

[g]Each hypothesis test involves a set risk of a Type I error. If a researcher searches or "fishes" through their data, testing many different hypotheses to find a significant effect, they inflate their Type I error rate.

[h]If the dependent and/or independent variables are not measured reliably, incorrect conclusions can be drawn.

[i]Restriction of range, such as floor and ceiling effects, reduce the power of the experiment, and increase the chance of a Type II error, because correlations are weakened by superficially reduced variability.

Additional files

Additional file 1: Criteria for judging the risk of bias in a study using the Cochrane Collaboration's 'Risk of bias' assessment tool (2011 version). Source Higgins et al. [66].

Additional file 2: Example of a 'Risk of bias summary' figure, presenting all of the judgements in a cross-tabulation of study by entry. Plus signs and the green colour represent a low risk of bias; minus signs and

the red colour represent a high risk of bias; question marks and the yellow colour represent an unclear risk of bias. Source: Higgins et al. [66].

Additional file 3: Further guidelines for a GRADE assessment: Going from assessments of risk of bias to judgements about study limitations for main outcomes. Source: Schünemann et al. [45].

Additional file 4: Worked example of quality assessment, using Environmental-Risk of Bias and Environmental-GRADE, on a study by Browne et al. (2000) [71].

Additional file 5: Worked example of quality assessment, using Environmental-Risk of Bias and Environmental-GRADE, on a study by Peach et al. (2001) [72].

Additional file 6: Worked example of quality assessment, using Environmental-Risk of Bias and Environmental-GRADE, on a study by Bradbury et al. (2004) [73].

Competing interests
The authors declare that they have no competing interests.

Authors' contribution
GB led the research for this article and drafted the manuscript in discussion with AM and IB, who provided comments throughout the writing process. All authors read and approved the final manuscript.

Authors' information
Dr. Gary Bilotta: Dr. Bilotta is currently a Senior Lecturer in Physical Geography at the University of Brighton. His research links the disciplines of hydrology, geomorphology, biogeochemistry and freshwater ecology. He has particular expertise in developing water quality monitoring and assessment tools for international water resource legislation. At the time of writing this article Dr Bilotta was seconded to the UK Department for Environment, Food and Rural Affairs (Defra), as part of the Natural Environment Research Council (NERC) Knowledge Exchange Scheme. The aim of this secondment was to drive improvements in the policy evidence-base and its use.
Dr. Alice Milner: Dr. Milner is currently a Research Associate at University College London specialising in climatic and environmental change on a range of time scales, past, present and future. Her research is focussed on understanding how ecosystem respond to abrupt climate changes during previous warm intervals, using terrestrial and marine sediments as natural archives of environmental change. At the time of writing, Dr. Milner was seconded to the UK Defra as part of the NERC Knowledge Exchange Scheme to improve the use and management of evidence in policy.
Prof. Ian Boyd: Professor Boyd is currently a Professor in Biology at the University of St Andrews and Chief Scientific Adviser to the UK Defra. His career has evolved from physiological ecologist with the NERC's Institute of Terrestrial Ecology, to a Science Programme Director with the British Antarctic Survey, Director at the NERC's Sea Mammal Research Unit, Chief Scientist to the Behavioural Response Study for the US-Navy, Director for the Scottish Oceans Institute and acting Director and Chairman with the Marine Alliance for Science and Technology for Scotland. In parallel to his formal positions he has chaired, co-chaired or directed international scientific assessments; his activities focusing upon the management of human impacts on the marine environment.

Acknowledgements
This work arises from Natural Environment Research Council funding (Grant *NE/L00836X/1*) in collaboration with the United Kingdom's Department for Environment, Food and Rural Affairs.

Author details
[1]University of Brighton, Brighton BN2 4GJ, UK. [2]Department for Environment, Food and Rural Affairs, London SW1P 3JR, UK. [3]University College London, London WC1E 6BT, UK. [4]University of St Andrews, Fife KY16 9AJ, UK.

Received: 8 April 2014 Accepted: 1 July 2014
Published: 1 September 2014

References
1. Begley CG, Ellis LM: **Drug development: raise standards for preclinical cancer research.** *Nature* 2012, **483**(7391):531–533.
2. Prinz F, Schlange T, Asadullah K: **Believe it or not: how much can we rely on published data on potential drug targets?** *Nat Rev Drug Discov* 2011, **10**(9):712–712.
3. Bohannan J: **Who's afraid of peer review?** *Science* 2013, **342:**60–65.
4. Godlee F, Gale CR, Martyn CN: **Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial.** *Jama* 1998, **280**(3):237–240.
5. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, Knipschild PG: **The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus.** *J Clin Epidemiol* 1998, **51**(12):1235–1241.
6. Centre for Reviews and Dissemination: **Systematic reviews: CRD's guidance for undertaking reviews on health care.** 2009, http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf.
7. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Sterne JA: **The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.** *Br Med J* 2011, **343:**d5928.
8. Gluud LL: **Bias in clinical intervention research.** *Am J Epidemiol* 2006, **163**(6):493–501.
9. Cozby PC: *Methods in behavioural research*. 10th edition. Boston: McGraw-Hill Higher Education; 2009.
10. Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, Fidler F, Rumpff L, Twardy C: **Expert status and performance.** *PLoS One* 2011, **6**(7):e22998.
11. Cooper H, Ribble RG: **Influences on the outcome of literature searches for integrative research reviews.** *Knowledge* 1989, **10:**179–201.
12. Oxman AD, Guyatt GH: **The science of reviewing research.** *Ann N Y Acad Sci* 1993, **703:**125–133.
13. Wells K, Littell JH: **Study quality assessment in systematic reviews of research on intervention effects.** *Res Soc Work Pract* 2009, **19**(1):52–62.
14. Moher D, Jadad AR, Tugwell P: **Assessing the quality of randomized controlled trials. Current issues and future directions.** *Int J Technol Assess Health Care* 1996, **12:**195–208.
15. Moher D, Pham B, Jones A, Cook DJ, Jadad AR: **Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?** *Lancet* 1998, **352:**609–613.
16. Colle F, Rannou F, Revel M, Fermanian J, Poiraudeau S: **Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain.** *Arch Phys Med Rehabil* 2002, **83**(12):1745–1752.
17. Herbison P, Hay-Smith J, Gillespie WJ: **Adjustment of meta-analyses on the basis of quality scores should be abandoned.** *J Clin Epidemiol* 2006, **59**(12):1249–1256.
18. Juni P, Witschi A, Bloch R, Egger M: **The hazards of scoring the quality of clinical trials for meta-analysis.** *J Am Med Assoc* 1999, **282**(11):1054–1060. 1999.
19. Deeks JJ, Dinnes J, D'amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DJ: **Evaluating non-randomised intervention studies.** *Health Technol Assess* 2003, **7**(27):1–179.
20. Collaboration for Environmental Evidence: **Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2.** 2013, www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf.
21. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S: **Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists.** *Control Clinical Trials* 1995, **16:**62–73.
22. Jüni P, Altman DG, Egger M: **Systematic reviews in health care: assessing the quality of controlled clinical trials.** *Br Med J* 2001, **323:**42.
23. Sterne JAC, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M: **Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research.** *Stat Med* 2002, **21:**1513–1524.
24. de Oliveira IR, Dardennes RM, Amorim ES, Diquet B, de Sena EP, Moreira EC: **Is there a relationship between antipsychotic blood levels and their clinical efficacy? An analysis of studies design and methodology.** *Fundam Clin Pharmacol* 1995, **9:**488–502.
25. Coleridge-Smith P: **The management of chronic venous disorders of the leg: an evidence-based report of an International Task Force.** *Phlebology* 1999, **14:**3–19.

26. Downs SH, Black N: **The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions.** *J Epidemiol Commun Health* 1998, **52:**377–384.
27. Lundh A, Gøtzsche PC: **Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies.** *BMC Med Res Methodol* 2008, **8**(1):22.
28. Greenland S, O'Rourke K: **On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions.** *Biostat* 2001, **2:**463–471.
29. Kunz R, Vist G, Oxman AD: **Randomisation to protect against selection bias in healthcare trials**. In *The Cochrane Library, issue 4.* Chichester: John Wiley and Sons Ltd; 2002. http://www.cochrane.org/reviews/clibintro.htm.
30. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Styliani I, Kokori MG, Tektonidou DG, Contopoulos-Ioannnidi DG, Lau J: **Comparison of evidence of treatment effects in randomized and nonrandomized studies.** *JAMA* 2001, **286:**821–830.
31. Als-Nielsen B, Gluud L, Gluud C: *Methodological quality and treatment effects in randomised trials - a review of six empirical studies.* Ottawa, Ontario, Canada: Canadian Cochrane Network and Centre, University of Ottawa; 2004. http://www.cochrane.org.
32. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, Bhandari M, Guyatt GH: **Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials.** *J Am Med Assoc* 2001, **285**(15):2000–2003.
33. Schulz KF, Chalmers I, Hayes RJ, Altman DG: **Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials.** *J Am Med Assoc* 1995, **273:**408–412.
34. Kjaergard LL, Villumsen J, Gluud C: **Reported methodological quality and discrepancies between large and small randomized trials in meta-analyses.** *Ann Intern Med* 2001, **135:**982–989.
35. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, Lau J: **Correlation of quality measures with estimates of treatment effect in metaanalyses of randomized controlled trials.** *J Am Med Assoc* 2002, **287**(22):2973–2982.
36. Als-Nielsen B, Chen W, Gluud L, Siersma V, Hilden J, Gluud C: *Are trial size and reported methodological quality associated with treatment effects? Observational study of 523 randomised trials.* Ottawa, Canada: 12th International Cochrane Colloquium; 2004.
37. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJG, Sterne JAC: **Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study.** *Br Med J* 2008, **336:**601–605.
38. Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR: **Publication bias in clinical research.** *Lancet* 1991, **337**(8746):867–872.
39. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG: **Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles.** *J Am Med Assoc* 2004, **291**(20):2457–2465.
40. Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Seida JK, Klassen TP: **Risk of bias versus quality assessment of randomised controlled trials: cross sectional study.** *Br Med J* 2009, **339:**b4012.
41. GRADE Working Group: **Grading quality of evidence and strength of recommendations.** *Br Med J* 2004, **328:**1490–1494.
42. Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan KL, Krishnan JA, Manthous CA, Maurer JR, McNicholas WT, Oxman AD, Rubenfeld G, Turino GM, Guyatt G: **An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations.** *Am J Respir Crit Care Med* 2006, **174:**605–614.
43. Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schünemann HJ: **Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians Task Force.** *Chest* 2006, **129:**174–181.
44. Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schünemann HJ: **An emerging consensus on grading recommendations?** *ACP J Club* 2006, **144:**A8–A9.
45. Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH: **Chapter 12: Interpreting results and drawing conclusions**. In *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011).* Edited by Higgins JPT, Green S. 2011. www.cochrane-handbook.org.
46. Atkins D, Best D, Briss PA, Eccles M, Falck Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW Jr, Zaza S: **Grade Working Group: Grading quality of evidence and strength of recommendations.** *Br Med J* 2004, **328**(7454):1490.
47. Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT, Hill S, Jaeschke R, Liberati A, Magrini N, Mason J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Vist GE, Williams JW: **GRADE Working Group: Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system.** *BMC Health Serv Res* 2005, **5:**25.
48. Stewart GB, Coles CF, Pullin AS: **Does burning degrade blanket bog.** *Systematic Review* 2004, **1:**1–29.
49. Roberts PD, Pullin AS: **Effectiveness of the control of ragwort (*Senecio*) species: Are currently recommended herbicides effective for control of ragwort (*Senecio*) species?** *CEE review* 2004, **04–003**. www.environmentalevidence.org/SR5a.html.
50. Felton A, Knight E, Wood J, Zammit C, Lindenmayer D: **Do tree plantations support higher species richness and abundance than pasture lands?** *CEE Review* 2010, **09–012**. www.environmentalevidence.org/SR73.html.
51. McDonald MA, McLaren KP, Newton AC: **What are the mechanisms of regeneration post-disturbance in tropical dry forest?** *CEE Review* 2010, **07–013**. www.environmentalevidence.org/SR37.html.
52. Stacey CJ, Springer AE, Stevens LE: **Have spring restoration projects in the southwestern United States been effective in restoring hydrology, geomporphology, and invertebrates and plant species composition comparable to natural springs with minimal anthropogenic disturbance?** *CEE Review* 2011, **10–002**. www.environmentalevidence.org/SR87.html.
53. Humbert J-Y, Pellet J, Buri P, Arlettaz R: **Does delaying the first mowing date increase biodiversity in European farmland meadows?** *CEE Review* 2012, **09–011**. www.environmentalevidence.org/SR72.html.
54. Pullin AS, Knight TM: **Support for decision making in conservation practice: an evidence-based approach.** *J Nat Conserv* 2003, **11:**83–90.
55. Hannan EL: **Randomized clinical trials and observational studies: Guidelines for assessing respective strengths and limitations.** *J Am Coll Cardiol Intv* 2008, **1:**211–217.
56. Stewart GB, Coles CF, Pullin AS: **Does burning of UK sub-montane, dry dwarf-shrub heath maintain vegetation diversity?** *CEE Review* 2004, **03–002**. www.environmentalevidence.org/SR2.html.
57. Stewart GB, Pullin AS, Coles CF: **Effects of wind turbines on bird abundance.** *CEE Review* 2005, **04–002**. www.environmentalevidence.org/SR4.html.
58. Stewart GB, Tyler C, Pullin AS: **Effectiveness of current methods for the control of bracken (*Pteridium aquilinum*).** *CEE Review* 2005, **04-001**. www.environmentalevidence.org/SR3.html.
59. Tyler C, Pullin AS: **Do commonly used interventions effectively control *Rhododendron ponticum*?** *CEE Review* 2005, **04–005**. www.environmentalevidence.org/SR6.html.
60. Tyler C, Clark E, Pullin AS: **Do management interventions effectively reduce or eradicate populations of the American Mink, *Mustela vison*?** *CEE review* 2005, **04–006**. www.environmentalevidence.org/SR7.html.
61. Roberts PD, Pullin AS: **Effectiveness of the control of ragwort (*Senecio*) species: Can biological control by the use of natural enemies effectively control *Senecio jacobaea* (common ragwort)?** *CEE Review* 2005, **04–004**. www.environmentalevidence.org/SR5b.html.
62. Roberts PD, Pullin AS: **The effectiveness of management options used for the control of *Spartina* species.** *CEE Review* 2006, **06–001**. www.environmentalevidence.org/SR22.html.
63. Roberts PD, Pullin AS: **The effectiveness of land-based schemes (incl. agri-environment) at conserving farmland bird densities within the UK.** *CEE Review* 2007, **05–005**. www.environmentalevidence.org/SR11.html.
64. Kabat TJ, Stewart GB, Pullin AS: **Are Japanese knotweed (Fallopia japonica) control and eradication interventions effective?** *CEE Review* 2006, **05–015**. www.environmentalevidence.org/SR21.html.
65. Ramstead KM, Allen JA, Springer AE: **Have wet meadow restoration projects in the Southwestern U.S. been effective in restoring hydrology, geomorphology, soils, and plant species composition to conditions comparable to wet meadows with minimal human-induced disturbance?** *CEE Review* 2012, **09–014**. www.environmentalevidence.org/SR75.html.
66. Higgins JPT, Altman DG, Sterne JAC: **Assessing risk of bias in included studies**. In *Chapter 8: Cochrane Handbook for Systematic Reviews of*

Interventions Version 5.1.0 (updated March 2011). Edited by Higgins JPT, Green S.: The Cochrane Collaboration; 2011. www.cochrane-handbook.org.
67. Chan AW, Altman DG: **Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors.** *Br Med J* 2005, **330**(7494):753.
68. Treasure T, MacCrae KD: **Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does.** *Br Med J* 1998, **317**(7155):362–363.
69. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norrisi S, Guyatt GH: **GRADE guidelines: 3. Rating the quality of evidence.** *J Clin Epidemiol* 2011, **64**(4):401–406.
70. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ: **GRADE: an emerging consensus on rating quality of evidence and strength of recommendations.** *Br Med J* 2008, **336:**924–926.
71. Browne S, Vickery J, Chamberlain D: **Densities and population estimates of breeding Skylarks Alauda arvensis in Britain in 1997.** *Bird Study* 2000, **47:**52–65.
72. Peach WJ, Lovett LJ, Wotton SR, Jeffs C: **Countryside stewardship delivers cirl buntings (Emberiza cirlus) in Devon, UK.** *Biological Conservation* 2001, **101:**361–373.
73. Bradbury RB, Browne SJ, Stevens DK, Aebischer NJ: **Five-year evaluation of the impact of the Arable Stewardship Pilot Scheme on birds.** *Ibis* 2004, **146:**171–180.

doi:10.1186/2047-2382-3-14
Cite this article as: Bilotta *et al.*: Quality assessment tools for evidence from environmental science. *Environmental Evidence* 2014 **3**:14.