Conservation Biology

Review

Reliability of evidence-review methods in restoration ecology

João P. Romanelli [1*] Paula Meli [2] Rafaela P. Naves [3] Marcelo C. Alves [4] and Ricardo R. Rodrigues [1]

[1]Laboratory of Ecology and Forest Restoration (LERF), ''Luiz de Queiroz" College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, Piracicaba, SP 13418–900, Brazil
[2]Laboratorio de Ecología del Paisaje y Conservación, Departamento de Ciencias Forestales, Universidad de La Frontera, Temuco, 4811230, Chile
[3]Department of Forest Sciences, ''Luiz de Queiroz" College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, Piracicaba, SP 13418–900, Brazil
[4]Informatics Technical Section, Luiz de Queiroz" College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, Piracicaba, SP 13418–900, Brazil

Abstract: In restoration science, evidence reviews play a crucial role in summarizing research findings in practice and policy. However, if unreliable or inappropriate methods are used to review evidence, decisions based on these reviews may not accurately reflect the available evidence base. To assess the current value of restoration reviews, we examined a sample of meta-analyses and narrative syntheses (n = 91) with the Collaboration for Environmental Evidence Synthesis Assessment Tool (CEESAT), which uses detailed criteria to assesses the method of policy-relevant evidence synthesis according to elements important for objectivity, transparency, and comprehensiveness. Overall, reviews scored low based on this standard: median score 16 out of 39, modal score 15, and mean 16.6. Meta-analyses scored higher than narrative syntheses (median 17 vs. 5, respectively), although there were some outlier narrative syntheses that had high scores, suggesting that quantitative synthesis does not solely reflect the reliability of a review. In general, criteria spanning the more fundamental review stages (i.e., searching for studies and including studies) received low scores for both synthesis types. Conversely, criteria comprising the later stages of the review (i.e., critical appraisal, data extraction, and data synthesis) were generally well described in meta-analyses; thus, these criteria achieved the highest individual CEESAT scores. We argue that restoration ecology is well positioned to advance so-called evidence-based restoration, but review authors should elucidate their conceptual understanding of evidence syntheses and recognize that conducting reliable reviews demands the same methodological rigor and reporting standards used in primary research. Given the potential of evidence reviews to inform management, policy, and research, it is of vital importance that the overall methodological reliability of restoration reviews be improved.

Keywords: Collaboration for Environmental Evidence, CEE, Collaboration for Environmental Evidence Synthesis Assessment Tool, CEESAT, evidence-based restoration, restoration synthesis, review reliability

Confiabilidad de los Métodos de Revisión de Evidencias en la Ecología de Restauración

Resumen: Para las ciencias de la restauración, las revisiones de evidencias juegan un papel muy importante en la síntesis de los hallazgos de las investigaciones en la práctica y en las políticas. Sin embargo, si se usan métodos poco fiables o inapropiados para revisar las evidencias, las decisiones que se tomen con base en estas revisiones pueden no reflejar acertadamente la base disponible de evidencias. Para analizar el valor actual de las revisiones de restauraciones examinamos una muestra de metaanálisis y síntesis narrativas ($n = 91$) con la Herramienta para la Síntesis de Análisis de la Colaboración para la Evidencia Ambiental (CEESAT), la cual usa criterios detallados para analizar el método de síntesis de evidencias relevantes para las políticas de acuerdo con los elementos importantes para la objetividad, transparencia y exhaustividad. En general, las revisiones tuvieron puntajes bajos con base en

*email joaopromanelli@hotmail.com
***Article Impact Statement**: Rigorous restoration reviews elucidate understanding of evidence syntheses and apply method and reporting standards of primary research.*
Paper submitted April 15, 2020; revised manuscript accepted September 10, 2020.

Conservation Biology, Volume 00, No. 0, 1–13

DOI: 10.1111/cobi.13661

este estándar (puntaje medio: 16 de 39, puntaje modal: 15, media: 16.6). Los metaanálisis tuvieron un puntaje más alto que las síntesis narrativas (mediana: 17 vs 5, respectivamente), aunque hubo algunas síntesis narrativas atípicas que tuvieron puntajes altos, lo que sugiere que la síntesis cuantitativa no refleja por sí sola la confiabilidad de una revisión. En suma, los criterios que abarcaron las etapas de revisión más fundamentales (es decir, buscar estudios e incluir estudios) recibieron puntajes bajos para ambos tipos de síntesis. Al contrario, los criterios que comprendieron las etapas tardías de la revisión (es decir, la valoración crítica, la extracción de datos y la síntesis de los datos) estuvieron generalmente bien descritos en los metaanálisis; por lo tanto, estos criterios alcanzaron los puntajes CEESAT individuales más altos. Argumentamos que la ecología de restauración se encuentra bien posicionada para adelantar la llamada restauración basada en evidencias, pero los autores de las revisiones deberían aclarar su entendimiento conceptual de la síntesis de evidencias y reconocer que la realización de revisiones confiables requiere el mismo rigor metodológico y los mismos estándares de reporte usados en la investigación primaria. Dado el potencial que tienen las revisiones de evidencias para orientar el manejo, las políticas y la investigación, es de vital importancia que se mejore la confiabilidad metodológica generalizada de las revisiones de restauración.

Palabras Clave: CEE, CEESAT, Colaboración para la Evidencia Ambiental, confiabilidad de revisión, Herramienta para la Síntesis de Análisis de la Colaboración para la Evidencia Ambiental, restauración basada en evidencias, síntesis de restauración

Introduction

Pervasive global environmental changes in the Anthropocene (Cooke et al. 2018) have been driving increasing challenges for restoration ecology and conservation science (McEuen & Styles 2019). Countries are relying on multiple international and regional policy initiatives to address the problems of rapid land degradation (Holl 2017; Romijn et al. 2019). The Bonn Challenge initiative, for example, arose as a global effort to restore 150 million ha of deforested and degraded lands by 2020 and 350 million ha by 2030 (Dave et al. 2019). The Initiative 20 × 20 has also arisen as a response to the Bonn Challenge (Romijn et al. 2019), aiming to restore over 50 million ha of degraded land by 2020 in Latin America and the Caribbean (WRI 2018).

To achieve such ambitious goals, researchers and decision makers need scientific guiding principles to optimize their efforts and resources (which are generally scarce) for large-scale projects (Holl 2017). Thus, with the increasing acknowledgment that scientific evidence reviews should be part of sound management and policy (Roberts et al. 2006; Philibert et al. 2012; O'Leary et al. 2016), specifically designed and reliable approaches are necessary. One way to achieve this is by advancing with so-called evidence-based restoration (Cooke et al. 2018).

Grown out of the broader evidence-based conservation and environmental management movement (Sutherland et al. 2004), and spearheaded by the Collaboration for Environmental Evidence (CEE), this concept comprises the application of rigorous, repeatable, and transparent methods to identify and amass relevant knowledge sources, critically evaluate the primary research, and produce reliable evidence syntheses to support any management intervention or activity, and it involves systematic reviews (SRs) (Cooke et al. 2018).

Systematic reviews, in contrast to a traditional narrative review (Table 1), have several methodological steps to ensure syntheses are reliable (CEE 2013; Vetter et al. 2013). These steps include the publication of an a priori protocol; comprehensive tried-and-tested searches across multiple bibliographic sources, including the gray literature; detailed information on the search and screen processes; critical appraisal of primary research; consistent extraction of data (descriptive information, metadata, quantitative or qualitative study findings); accurate synthesis of study findings through appropriate quantitative (e.g., meta-analysis) or qualitative (e.g., meta-ethnography) methods; and full documentation of all review activities to allow verification and repeatability of methods by a third party (Haddaway et al. 2015).

Methods for conducting SRs are now the gold standard of evidence synthesis (Pullin & Stewart 2006), and since they were introduced to ecology nearly 2 decades ago (Grames & Elphick 2020), their use has increased considerably (Haddaway et al. 2015). However, although many current evidence syntheses are carried out by science-based organizations (e.g., government or nongovernmental organizations) and by academics, this does not necessarily mean that contemporary restoration evidence syntheses are based on the most rigorous methods and available evidence (Cooke et al. 2018). Instead, there is sufficient evidence to support the claim that review requirements are perceived differently across different types of synthesis (e.g., traditional reviews vs. meta-analyses) (Roberts et al. 2006; O'Leary et al. 2016), and that even in the most prestigious journals the quality of science can be highly variable (Smith 2006; Alpert 2007; Cooke et al. 2018).

Admittedly, restoration ecology is still a young discipline (Romanelli et al. 2018), but outcomes in ecological restoration often change quickly over time (Holl

Table 1. Evidence synthesis and review terminology.

Term	*Definition*	*Reference*
Scientific evidence	information gathered from scientific research through a scientific method that derives repeatable and reproducible findings countering or supporting a hypothesis or theory	Berger-Tal et al. 2018
Evidence review	overarching term for articles that collate and summarize multiple primary studies related to a specific, policy-relevant question	CEE 2013, 2018
Evidence synthesis	process of gathering information from a range of sources to inform decisions on specific issues; "occurs once the evidence base has been accumulated and the data of interest extracted"	O'Leary et al. 2016; CEE 2018
Narrative synthesis	process in which prose is used to summarize and draw conclusions from primary research and that may be supplemented by reviewers' own experiences; some include limited quantitative analyses	O'Leary et al. 2016
Systematic review	"review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review. statistical methods (meta-analysis) may or may not be used to analyze and summarise the results of the included studies"	CEE 2013, 2018; Berger-Tal et al. 2018
Systematic map	collation, description, and cataloguing of available evidence relating to a topic or question of interest; included studies can be used to identify evidence for policy-relevant questions, knowledge gaps (to help direct primary research), and knowledge clusters (subsets of evidence that may be suitable for secondary research, e.g., systematic review); does not attempt to answer a specific question, as for systematic reviews	CEE 2013, 2018, Berger-Tal et al. 2018
Protocol	document produced prior to the commencement of evidence synthesis; describes background to the synthesis, questions, strategy to be used to search for primary research articles, and criteria for deciding whether or not an article is then relevant to include in the synthesis; outlines approach to assessing the quality of each included study and to extracting and synthesizing data from primary research articles (CEE, 2013); analogous to developing and documenting a method prior to conducting fieldwork or experiments and is similarly integral to producing a study that is robust to post hoc changes in methods and scope	CEE 2013; Woodcock et al. 2014
Meta-analysis	"set of statistical methods for combining the magnitude of the outcomes (effect sizes) across different data sets addressing the same research question"	Koricheva et al. 2013; Berger-Tal et al. 2018

2017). Consequently, if restoration practitioners are failing to make full use of the available body of evidence or account for study quality—both of which are common pitfalls with traditional reviews (Roberts et al. 2006; O'Leary et al. 2016; Berger-Tal et al. 2018)—there is much room for the implementation of actions that are unsound and potentially counterproductive (Lajeunesse & Forbes 2003; Cooke et al. 2018).

For instance, several recent meta-analyses in restoration ecology aimed to evaluate whether natural regeneration is more effective than active restoration at recovering tropical forests. Reid et al. (2018) found that some comparisons between strategies were biased by positive site selection, which led reviews to diverge in their conclusions, casting doubt on the robustness of methods used by studies to appraise evidence. Likewise, controversial reviews were reported in conservation science (Grames & Elphick 2020) in which the methods reported in the studies were not reproducible and focused on opaque decisions about which evidence to include and which to omit that could change review conclusions.

Accordingly, with so many reviews on restoration topics being published (Appendix S1), it is valuable to have an overview of their reliability that highlights both strengths and opportunities for improvement. Thus, assessment tools with which individual reviews can be evaluated to determine their value are of vital importance. Within environmental science, an assessment tool expressly intended for evaluating environmental evidence reviews is the Collaboration for Environmental Evidence Assessment Tool (CEESAT) (Woodcock et al. 2014). The CEESAT is based on environmental SR methods and is transferable to all types of reviews that use literature review techniques (CEE 2013; O'Leary et al. 2016).

We applied CEESAT to evaluate the reliability of restoration reviews to enable researchers and decision makers to identify reliable and unbiased reviews. For clarity, the evidence-review terms we used are presented in Table 1. We specifically assessed the reliability of meta-analyses and narrative syntheses based on information reported in each study; the relationship between reliability scores and synthesis types; and the relationship between Scimago journal rank (SJR) and reliability scores. We also considered misuses of review terminology across restoration ecology and the implications of reliability scores

for restoration and conservation decision making. We compiled a list of critical steps to advance evidence-based restoration.

Methods

Literature Search

Restoration ecology is the scientific field that supports the practice of ecological restoration (Aradottir & Hagen 2013; Romanelli et al. 2018); therefore, these 2 terms ("*ecological restoration" OR "restoration ecology*") were used to retrieve titles, abstracts, and keywords of related publications. We combined these terms with *"systematic* review*" OR "meta-analys*" OR "meta analys*" OR "metaanalys*" OR "metanalys*"* to define our population of reviews (Appendix S1).

We acknowledge that we could have underestimated the total number of all available evidence syntheses in restoration ecology because documents described with other terms (e.g., *forest restoration* or *ecosystem restoration*) may also be relevant (Guan et al. 2018). Additionally, our search strategy depended partly on how reviews were described by their authors. Nonetheless, SRs (which may or may not include meta-analyses [Table 1]) should follow rigorous guidelines expressly designed to ensure objectivity and transparency in review methods (Pullin & Stewart 2006; Moher et al. 2009) so as to provide an appropriate basis for assessing reliability (Woodcock et al. 2014). If a review does not follow these guidelines, then, it is not an SR, irrespective of how the authors have designated their work. By using this search strategy, we sought to provide a platform in which to investigate potential misuses of review terminology in restoration ecology, given that several studies (e.g., Côté & Reynolds 2012; Vetter et al. 2013; Koricheva & Gurevitch 2014; Gurevitch et al. 2018; Reid et al. 2018; Grames & Elphick 2020) have reported inconsistencies related to methods and usage of review terminology in ecology and conservation. Regardless of previous conceptual analyses, we posteriorly classified reviews as either meta-analysis or narrative synthesis (Table 1), so that we could perform a reliability assessment based on the type of synthesis conducted in each study (see "Reliability Assessment and Data Analyses").

Bibliographic Sources

We selected 4 widely used bibliographic sources in the fields of conservation and environmental science to test the reliability of restoration reviews: Web of Science (core collection: SCI-E, SSCI, and ESCI), Scopus, CAB Direct, and SciELO. We also used Google Scholar as a search engine. We automated searches in Google Scholar with Publish or Perish (PoP) (http://www.harzing.com/pop.htm) freeware, which places results in several formats. Data were retrieved on 23 October 2019 and were last updated on 31 October 2019 (Appendix S1).

Selection of Studies

All retrieved articles were screened for relevance. Titles and abstracts were screened according to the following inclusion criteria: population, reviews should be a synthesis of primary research, described by authors (in title, abstract, or keywords) as an SR or a meta-analysis and intervention, and reviews should address the ecological restoration of terrestrial ecosystems (any outcome was eligible for inclusion).

Recognizing the potential for subjective decisions in this step (inclusion-exclusion decisions) (CEE 2018), all articles retrieved after title, abstract, and keyword screening ($n = 127$) were also analyzed by a second and third assessor. Assessor decisions were compared using the kappa test of agreement (Cohen 1960). Kappa scores of 0.94 and 0.96 (95% lower and upper confidence limits $n = 127$) were obtained, which indicated almost perfect agreement between assessors and that decisions were sufficiently repeatable (CEE 2018). To demonstrate transparency in our methods (CEE 2018), a list of all articles excluded from the analysis, based on full-text assessment, and the reasons for exclusion are provided in Appendix S2.

Reliability Assessment and Data Analyses

Selected publications were scored according to the CEESAT (Woodcock et al. 2014). The CEESAT consists of a set of 13 criteria that align with environmental SR method (CEE 2013) (Table 2). These criteria are used to evaluate the reliability of reviews in terms of objectivity, transparency, and comprehensiveness. *Reliability* refers to the level of confidence that an end user may place in some specific review method, not in the accuracy of its results (O'Leary et al. 2016). We followed the explanatory guidelines produced by Woodcock et al. (2014) and simplified by O'Leary et al. (2016). For each criterion, we determined whether optimal (3 points), intermediate (1 point), or inadequate standards (0 points) were applied based on how well the criterion was met (maximum of 39 points in the final score). Reviews that applied meta-analytical techniques scored 3 points in CEESAT (criterion 6.1 [Table 2]); narrative syntheses in which some quantitative analysis was conducted (e.g., descriptive statistics) scored 1 point, and reviews that were solely narrative syntheses scored 0 points. A summary of the rationale for each criterion is provided in Table 2.

Reviews selected after the full-read text ($n = 91$) were scored by 1 scorer (J.P.R). However, before performing the full analyses, 2 samples containing 20 random

Table 2. Summary of Collaboration for Environmental Assessment Tool (CEESAT) criteria for scoring reliability of reviews (Woodcock et al. 2014; O'Leary et al. 2016).

Criterion	*Rationale*
1. Protocol was available for review or comment before the synthesis was conducted.	prevents post hoc changes in methods, objectives, and scope, thereby increasing robustness of review
2: Search for studies	
2.1 search for literature is based on a comprehensive range of sources	increases likelihood of capturing available evidence base and reduces publication bias
2.2 search strings are clearly defined	enables external evaluation, allows search to be repeated, and avoids open-ended searches
3. Including studies	
3.1 clearly documented inclusion criteria applied to all potentially relevant studies found during the search	reduces risk of subjective decisions regarding included studies; reduces selection bias
3.2 inclusion and exclusion decisions are repeatable	demonstrates objectivity of study in- and/exclusion decisions
3.3 inclusion and exclusion decisions are transparent	enables external verification of in- and exclusion decisions
4. Critical appraisal	
4.1 critical appraisals of methods are conducted and reported for each included study	assesses quality of evidence available for synthesis in terms of susceptibility to bias
4.2 included studies are objectively weighted according to methodological quality	ensures greater emphasis is given to more robust studies
5. Data extraction	
5.1 data extraction is documented, repeatable, and consistent	enables external validation and repetition and reduces the potential for bias
5.2 extracted data are reported for each study	enables external verification and analysis of extracted data
6: Data synthesis	
6.1 quantitative synthesis is conducted	increases objectivity by reducing potential for subjective assessment of findings
6.2 heterogeneity in the effect of the intervention or exposure is investigated statistically	indicates external and general applicability of results and appropriateness of combining studies
6.3 synthesis considers possible publication bias	assesses potential for publication bias to influence review findings

publications (approximately 22% of the data set) were scored by 2 other scorers (P.M. and R.P.N.) to account for possible differences in the application of scoring criteria and potential biases introduced by different scorers' expertise. All scorers were professionals in the broad field of environmental sciences.

We focused on the methodological reliability of restoration reviews as a whole to identify strengths and weaknesses across this population and offer guidelines for the improvement of future syntheses. Our goal was not to name and shame individual papers. Consequently, we anonymized the list of included reviews as well as the scores assigned.

Scoring decisions (between scorers 1 and 2 and between scorers 1 and 3) were analyzed considering the magnitude of disagreement between scorers with a weighted kappa test of agreement. Final kappa values ranged from 0 to 1, and higher kappa values indicated greater agreement (Cohen 1960; Landis & Koch 1977). We used descriptive statistics (median, mode, and mean) to enable comparisons between synthesis types. Differences in the mean scores between reviews were tested statistically with a Mann–Whitney U test for pairwise comparisons. Pearson's correlation coefficient was used to analyze the relationship between journal impact factor (SJR) and CEESAT scores. The Kruskal–Wallis one-way analysis of variance on ranks was used to detect differences among scores awarded for meta-analyses and narrative reviews. We used R to perform statistical analyses and produce figures (R Core Team 2019). We created network maps with the VOSviewer software (version 1.16.15) with the text-mining function to construct and visualize co-occurrences of keywords and examined trends across the included literature. VOSviewer is a software tool expressly designed for the analysis of bibliometric data (van Eck & Waltman 2010).

Results

Summary of Reviews

We identified 91 reviews that met our inclusion criteria, among which approximately 79% ($n = 72$) were meta-analyses and approximately 21% ($n = 19$) were narrative syntheses. We found 72 different bibliographic sources used by restoration authors to gather evidence from primary studies, among which the Web of Science was the most cited (approximately 74% of studies, $n = 67$), followed by Google Scholar (approximately 26%, $n = 24$) and Scopus (approximately 18%, $n = 17$). Cited databases (e.g., previously published reviews) were used as a bibliographic source for approximately 14% of studies ($n = 13$). Gray literature sources were specialist websites (approximately 10%; $n = 9$), theses, dissertations (approximately 9%; $n = 8$), and Google Scholar (as a

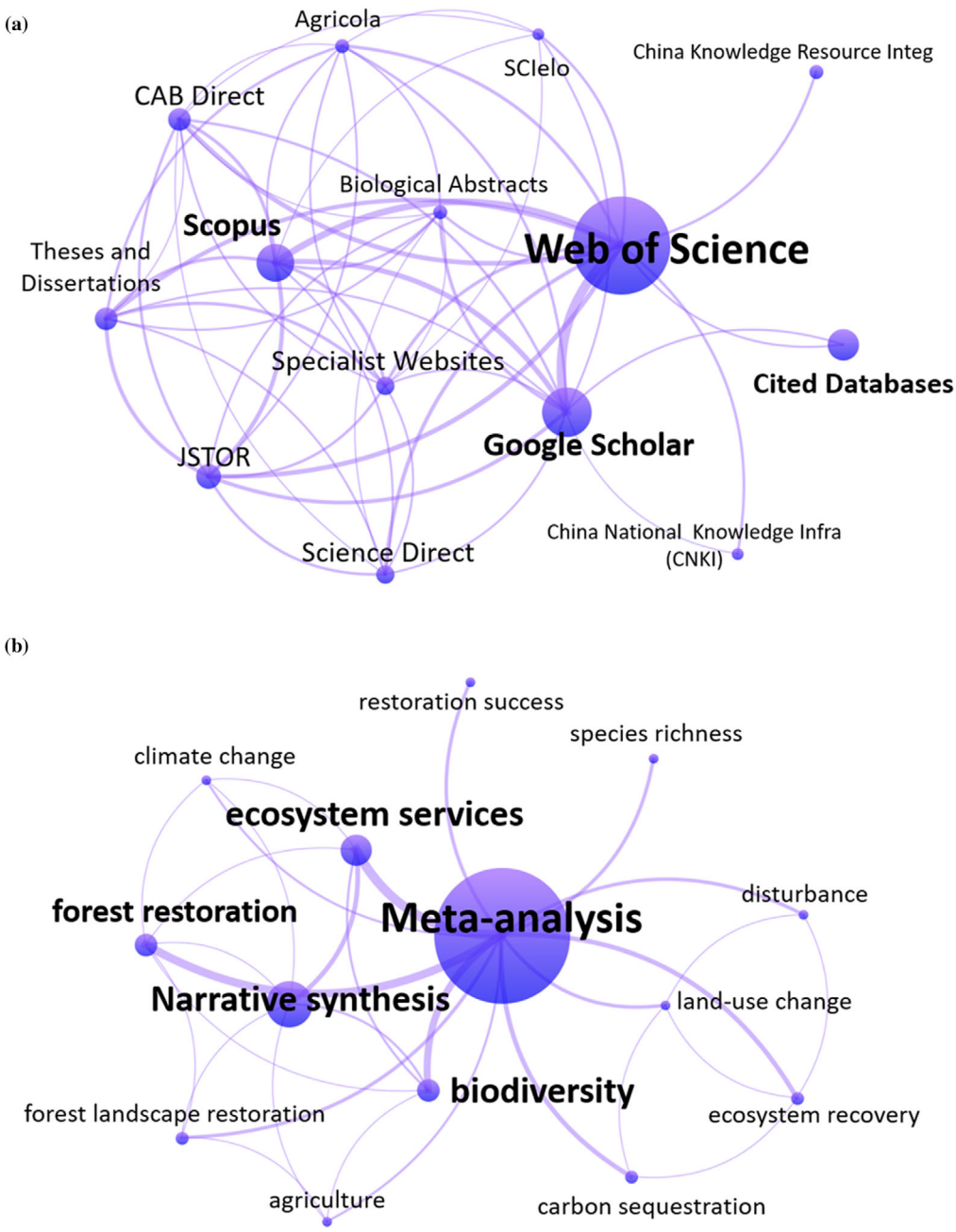

Figure 1. Results of network analysis of the (a) main bibliographic sources used by authors of restoration reviews to gather evidence from the primary literature and (b) main research topics addressed by reviews through the analysis of the author's keywords. Network maps are limited to present items with a minimum of 3 occurrences (the same term or expression). Size of the node is proportional to the number of occurrences, and thickness of the edges represents co-occurrences between items.

search engine) (Fig. 1a). Reviews covered both negative anthropogenic impacts (e.g., land-use change) and restoration and conservation strategies (e.g., forest landscape restoration and forest reserve). Topics related to ecosystem services, biodiversity, and forest ecosystems accounted for most of the occurrences across this body of literature (out of nearly 400 author supplied keywords used to describe reviews) (Fig. 1b).

Overall Reliability of Restoration Reviews

Individual reviews spanned the entire CEESAT score gradient (0–39) (Fig. 2a; Appendix S2). The total mean score (considering both meta-analyses and narrative syntheses) was 16.6 (mode = 15, median = 16). In general, the lowest mean scores were found among criteria spanning the earlier stages of the review process (i.e., publishing an a

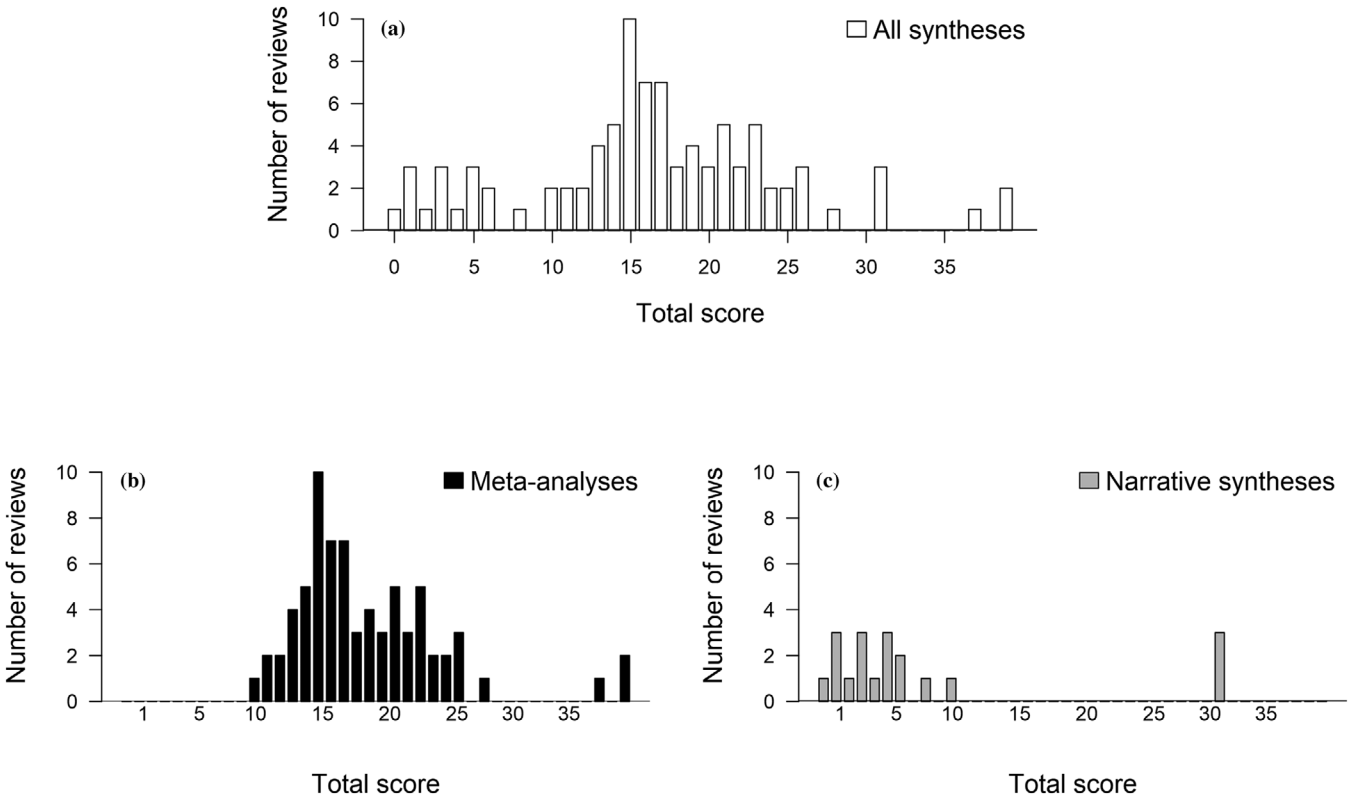

Figure 2. Distribution of restoration reviews by overall Collaboration for Environmental Assessment Tool (CEESAT) score assigned according to synthesis type: (a) all syntheses (n = 91), (b) meta-analyses (n = 72), and (c) narrative syntheses (n = 19). Total score is presented in 5 intervals to the maximum available score of 39.

priori protocol, searching for studies, and including studies) (Fig. 3). Except for criterion 3.1 (clearly documented inclusion criteria; mean score = 2.5), all other criteria comprising these more fundamental review stages had a mean score <0.8 (mode = 0). Conversely, criteria spanning the later stages of the evidence synthesis (i.e., critical appraisal of studies, data extracting, and data synthesis) achieved higher mean scores among CEESAT criteria (range 1.1–2.4).

The median total score was 17.0 for meta-analyses (mean score = 18.8) (Fig. 2b) and 5.0 for narrative syntheses (mean score = 8.1) (Fig. 2c); the differences were significant (Mann–Whitney U, $W = 207.5$, $p < 0.05$). Therefore, we evaluated these 2 groups separately, considering the effect of the synthesis type.

Reliability among Meta-Analyses and Narrative Syntheses

Reliability scores for meta-analyses and narrative reviews showed significantly different results ($n = 72$, $n = 19$; $p < 0.05$) for all criteria spanning the later stages of the review process and for criterion 3.2 (repeatability of inclusion/exclusion decisions) (Appendix S2). The mean score for all these criteria was higher for meta-analyses (range 1.3–3.0; mode range 0–3.0) than for narrative syntheses (mean score range 0–0.7; mode = 0).

Conversely, most criteria that comprise the initial review stages presented no statistically significant differences among synthesis types (Kruskal–Wallis one-way analysis of variance on ranks, $n = 72$, $n = 19$; $p > 0.05$). However, narrative syntheses showed a slight tendency to score higher than meta-analyses when authors published an a priori protocol before publication of the review (mean scores 0.5 and 0.2, respectively; mode = 0), when the search process was repeatable (criterion 2.2; mean score 0.7 vs. 0.5; modal values = 0), and when there was transparency in the inclusion-exclusion decisions (criterion 3.3, Appendix S2) (mean scores 0.6 and 0.2, respectively; mode = 0). In contrast, meta-analyses scored slightly higher than narrative syntheses when comprehensive searches were performed to locate relevant studies (criterion 2.1, Appendix S2) (mean scores 0.8 and 0.7, respectively; mode = 0) and when criteria for inclusion were defined precisely (criterion 3.1, Appendix S2) (mean scores 2.7 and 1.9, respectively; mode = 3); nonetheless, scores for both synthesis types remained low overall.

Repeatability Test

Agreement among assessors (between A 1 and 2 and between A 1 and 3) was generally high for most of the

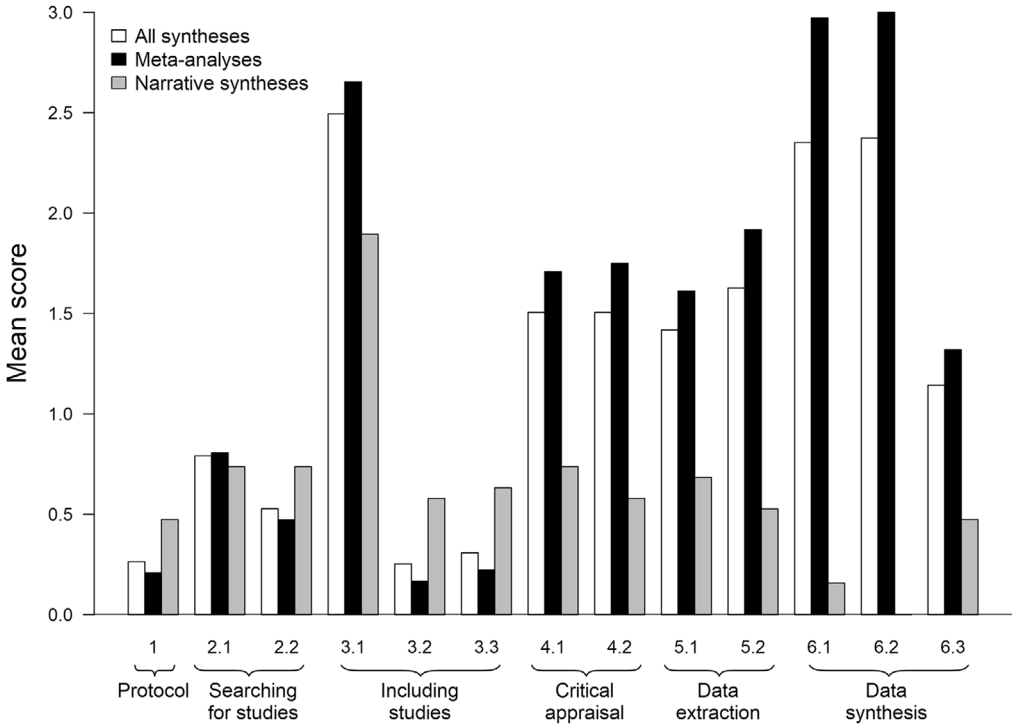

Figure 3. Mean scores for all restoration syntheses, meta-analyses, and narrative syntheses across Collaboration for Environmental Assessment Tool (CEESAT) criteria.

Table 3. Repeatability of scoring of individual Collaboration for Environmental Assessment Tool (CEESAT) criteria evaluated based on percent agreement between assessors (i.e., repeatability between scorers 1 and 2 and between scorers 1 and 3) and kappa statistic weighted according to the extent of disagreement.*.

	Agreement (%)		*Weighted kappa*	
CEESAT criterion	*A 1–2*	*A 1–3*	*A 1–2*	*A 1–3*
1. Protocol	100	100	1	1
2.1 Search sources	60	75	0.53	0.81
2.2 Search strings stated	35	30	0.09	0.16
3.1 Inclusion criteria documented	70	55	0.55	0.17
3.2 Inclusion decisions are repeatable?	90	100	0.61	1
3.3 Inclusion/exclusion decisions transparent?	90	95	0.61	0.82
4.1 Critical appraisal of methods	50	45	0.33	0.38
4.2 Objective weighting	50	60	0.32	0.60
5.1 Data extraction documented	45	60	0.35	0.51
5.2 Extracted data reported	60	80	0.42	0.78
6.1 Quantitative synthesis	90	75	0.92	0.82
6.2 Heterogeneity investigated	95	60	0.88	0.22
6.3 Consideration of publication bias	45	65	0.39	0.76

**Numbers represent the mean agreement across criteria for each combination of scorer 1 and the second scorer. Interpretation of kappa test: for example, a 0 versus 1 disagreement is less important than a 0 versus 3 disagreement.*

individual criteria (100% agreement for 2 criteria and at least 60% for the other 9 criteria) (Table 3). Weighted kappa test showed perfect agreement and substantial agreement between A 1 and 2 and between A 1 and 3, respectively. These results indicated that scores were consistent among scorers for most criteria. However, agreement for criteria 2.2 and 4.1 was lower (range 35–50%) (slight and fair agreement based on weighted kappa test).

SRJ Impact Factor and Review Reliability

There was no significant correlation between SRJ and CEESAT scores (considering both meta-analyses and narrative syntheses) (Pearson's $r = 0.0463$, $p = 0.663$, $n = 91$). Moreover, no significant correlation was found when meta-analyses (Pearson's $r = -0.0714$, $p = 0.551$, $n = 72$) or narrative syntheses (Pearson's $r = -0.0107$, $p = 0.965$, $n = 19$) were tested separately. However, 2

of 72 meta-analyses and 3 of 19 narrative reviews that achieved the highest CEESAT scores (>30) were published by the same journal (the CEE journal *Environmental Evidence*).

Discussion

Restoration researchers have been using several approaches and bibliographic sources (Fig. 1a) to locate relevant literature and synthesize meaningful information. Trends in research topics (Fig. 1b) were congruent with other recent studies in which conservation and restoration subjects are discussed side by side (Possingham et al. 2015; McMahen & Bommel 2020), and topics related to biodiversity conservation and ecosystem services are key issues for future integrated research in ecological restoration (Guan et al. 2018; Staples et al. 2019). Our population of reviews is, therefore, likely to be perceived as an important subset of publications in restoration ecology and to include topics of interest to researchers and policymakers.

Reliability of Restoration Reviews

Total CEESAT scores provide summary information that can be used to gauge the reliability of evidence syntheses (Woodcock et al. 2014). The distribution of total scores (Fig. 2) indicated that there were very good, good, average, and poor reviews published in restoration ecology. The greatest discrepancy in this distribution was among narrative syntheses, which were separated into 2 groups, very good and poor methodological reviews, where the latter was more common. Most meta-analyses, in turn, had scores from 10 to 25 and were assessed as average to good reviews, but some high-scoring meta-analyses were also found.

In addition to the total CEESAT scores, individual scores for each criterion can be very informative; they show strengths and specific aspects that need to be improved (O'Leary et al. 2016). Each criterion can be used to discriminate syntheses, and for most criteria, a spread of scores was found (Fig. 3). Although meta-analyses scored higher than narrative syntheses for most criteria, some individual criteria scored low for both review types, and some were followed to an extent, but could be improved. In accordance with previous studies (Roberts et al. 2006; O'Leary et al. 2016; Grames & Elphick 2020), we found that authors did not consistently apply systematic-review principles to all review stages, preventing high scores from being achieved for most reviews.

Overall, both review types scored low on publishing the protocol a priori; only 8 of 91 reviews reached the optimal score. This was expected because the majority of reviews did not use or cite CEE guidelines. Rather, we found that 9 of 91 reviews mentioned the PRISMA checklist as a workflow. Nonetheless, due to its development for guiding SRs in healthcare (Moher et al. 2009), PRISMA has limited applicability for reviews in conservation and environmental management (Haddaway et al. 2018). Curiously, this specific set of reviews (9 of 91) was published after 2014, when CEE guidelines were already available (CEE 2013). When a review protocol is mentioned in the synthesis and is robust, the credibility of the review increases (CEE 2018) because methods for evidence synthesis had been peer reviewed.

Restoration syntheses also failed when measured based on comprehensiveness (criterion 2.1, Fig. 3). Most reviews (approximately 51%) used only a single bibliographic source to retrieve information from primary research. However, reliable evidence syntheses must have comprehensive search strategies (Haddaway 2017). Because the entire review process is built on the results of this stage, searches must be sufficiently sensitive to gather as much of the evidence relevant to the review as possible (Bayliss & Beyer 2015). Moreover, a remarkably common lack of understanding about the Web of Science platform was evidenced. The WoS is not a single database, as some authors claimed; rather, it is a platform from which a range of databases can be queried (Haddaway 2017). Poor reporting of search and inclusion processes (i.e., criteria 2.2 and 3.3, Fig. 3) mean that the tenet of repeatability can be easily compromised (Moher et al. 2009; Grames & Elphick 2020)—something that is a central principle of evidence-based science (Parker et al. 2016).

Both review types scored relatively high on the inclusion criteria (criterion 3.1, Fig. 3). Nevertheless, this does not guarantee the author's conduct in selecting or including potentially relevant studies. Decisions over which studies are relevant for inclusion should be based on clearly defined criteria, and the decision process should be repeatable and transparent. In this regard, both review types failed to sufficiently demonstrate that their decisions are repeatable (criterion 3.2) and transparent (criterion 3.3), but narrative synthesis presented a slight tendency to report and better document these stages. These are, therefore, key steps in the early stages of the review process where the reliability of restoration syntheses should also be enhanced.

The problem of the variable quality in research can also affect the reliability of both qualitative and quantitative reviews considerably (Haddaway et al. 2015). Most restoration syntheses failed to report how primary studies were evaluated (i.e., the assessment of internal and external validity of each primary study) (criterion 4.1) and to document such activities (criterion 3.3—an imperative requirement for the full-text assessment in the including-studies stage) (CEE 2018). Although each piece of research may be relevant for inclusion in reviews, failing to appraise primary research and giving equal

weight to each study (criterion 4.2) can lead to misleading conclusions (Englund et al. 1999; Gates 2002), potentially overestimating (or underestimating) the effect of an intervention or exposure. In this respect, reviews fell short of optimal scores for criteria at the critical-appraisal stage, mainly for narrative syntheses. Although they were followed to some extent in meta-analyses, they could be improved.

Overall, full reporting and repeatability of the data extraction (criteria 5.1 and 5.2) led to average to good scores for meta-analyses and lower scores for narrative syntheses. Review authors frequently must make decisions on which results to extract and how any subsequent qualitative or quantitative analysis should be objectively explained and data effectively reported. Because the volume and type of data from primary evidence may vary considerably (Gates 2002), failing to minimize bias and establish consistent criteria to extract metrics in reviews may introduce ambiguity in the evidence synthesis if variables are interpreted differently. Consequently, methods are unreliable when they are unrepeatable. To improve the reliability of data extraction, review authors must follow protocols for coding studies and include operational definitions of variables; then, an independent end user can reassess methods employed and judge their reliability (Grames & Elphick 2020).

At the data synthesis stage, criteria 6.1 and 6.2 achieved the highest individual scores among all CEESAT criteria. This was mainly due to the number of meta-analyses in our sample ($n = 72$ of 91). Meta-analyses generally provided well-defined methods to demonstrate how primary studies were quantitatively synthesized (criterion 6.1) and how the heterogeneity of the intervention-exposure effects was investigated (criterion 6.2). Conversely, narrative syntheses did not perform well for these criteria. Although this difference is reflected in CEESAT criteria distinguishing between synthesis types (Woodcock et al. 2014), there is also considerable variation in each of these 2 categories of reviews, including some outlying narrative syntheses that had higher score than several meta-analyses.

Traditional reviews and even formal meta-analyses can be highly susceptible to bias at various stages of the review process (Haddaway et al. 2015; 2018). Often, authors only search for and include academic publications in their synthesis—where results with negative or no effect are less likely to be published—introducing publication bias into their methods (Jennions & Møller 2002; Jennions et al. 2013). Overall, authors of both types of syntheses failed to evaluate the robustness of their conclusions by assessing the likelihood of publication bias (criterion 6.3). Nonetheless, whereas meta-analyses achieved relatively average to low scores for this criterion, most narrative syntheses scored low. The majority of reviews showed no effort to address publication bias by using any of the broad classes of approaches available, or if they did, publication bias was considered subjectively. This is, therefore, another critical point to be perceived by restoration authors to improve the reliability of their syntheses. Some strategies that can be taken to counter such problems and possibly mitigate them include searching for gray literature, contacting people generating data in the field to access their data sets (Bayliss & Beyer 2015), and searching for studies in languages other than English (Leimu & Koricheva 2005).

Effect of Synthesis Type on Review Reliability

The large variation in scores for both meta-analyses and narrative syntheses suggests that the type of synthesis conducted in reviews should not be used as a conclusive indicator of methodological reliability.

Certain activities in the review process were found to lead to greater reliability. For example, meta-analyses scored significantly higher than narrative synthesis for 7 of 13 criteria. The higher scores achieved by meta-analyses are in part because certain criteria required statistical analyses to be assigned a high score (i.e., critical appraisal and data synthesis). However, points for these criteria can also be given to narrative syntheses (O'Leary et al. 2016). The relatively high scores found for McDonald et al. (2010), Bernes et al. (2015), and Gutiérrez Rodriguez et al. (2016) indicated that even when meta-analysis is not performed, many other aspects of reviews can be conducted robustly and are recognized as such by CEESAT. Importantly, while CEESAT partly reflects whether or not a meta-analysis was conducted, a quantitative synthesis does not solely reflect the reliability of all the methods employed. In other words, conducting a meta-analysis does not guarantee a rigorous conclusion if other steps in the review process are not adequately carried out (Harrison 2011).

Relationship between Impact Factor and Review Reliability

Overall, researchers cite earlier published studies to support, describe, or develop a particular point of view. The citation is, therefore, an indication of the importance that the scientific community attaches to the research (Okubo 1997). Moreover, citations are a measure of the overall impact of an article's influence, or that of its authors, and are also used as a variable to identify the most influential journals (Seglen 1992; Okubo 1997). Journal impact factor is sometimes used as an indicator of review quality, but this is neither equivalent nor unequivocally correlated with scientific quality (Seglen 1992). Nevertheless, we considered the correlation analysis between SJR and CEESAT scores, as proposed by O'Leary et al. (2016), mainly to inform nonspecialists and other restoration readers of the minimum caution that they must take when selecting more reliable reviews based on journal impact factor.

Although we observed a tendency of journals with higher impact factors to publish reviews with relatively high scores, this effect was not significant because we also found the opposite. The highest scores (>30) were assigned to papers (n = 5/91) published by the same journal, suggesting that the journal's policies differed from other journals in their format, guidelines, and underlying review process. According to Bradford's law, a skewed distribution in a journal's influence is expected in bibliometrics (Romanelli et al. 2018) because only a few journals are devoted to specific subjects, although many others may publish articles on that theme (Okubo 1997). This fact is easily seen when considering the transdisciplinary approach of meta-analyses or SRs (Diefenderfer et al. 2016), which are considered for publication across several journals. Overall, our results indicated that the reliability of restoration reviews should not be judged solely on the impact factor of the journal in which they are published.

Misuses of Review Terminology

We identified an imprecise and inaccurate use of review concepts (Table 1) among restoration reviews, mainly between narrative synthesis and SRs and meta-analysis and SRs. Several narrative syntheses in our sample claimed to be SRs (n = 16/19), but these syntheses did not attempt to follow SR guidelines (e.g., Pullin & Stewart 2006; Higgins & Green 2011; CEE 2018) that reduce error and bias; therefore, they should be conceptually classified among other narrative approaches rather than SRs. Likewise, we found that some meta-analyses were designated as SRs (n = 5/72), but the same problem arose because these syntheses did not consistently apply SR principles in all review stages. Indeed, these reviews failed even to address the more fundamental stages (e.g., publishing a protocol a prior or searching in multiple bibliographic sources). We also found 2 studies in which data were extracted from several primary studies, but the review essentially represented a qualitative synthesis rather than a meta-analysis (as assigned by their authors). Although this was not a major problem among reviews we evaluated, review authors should understand that, by definition, a meta-analysis involves a quantitative statistical synthesis combining outcomes (effect sizes) across different data sets addressing the same research question (Vetter et al. 2013; Gurevitch et al. 2018).

Implications for Decision Making

Given that evidence reviews are increasingly used to inform decisions with relevant environmental and socioeconomic implications, they need to be particularly reliable (Halme et al. 2010; Haddaway et al. 2015). Thus, when selecting reviews for decision making, we recommend that the review reliability be considered on a case-by-case basis, preferably through assessment with a standardized tool, such as CEESAT. Because environmental decisions are often made in limited time frames (O'Leary et al. 2016), we believe that CEESAT could be incorporated into the decision-making process given the quality of a review can be quickly accessed. For example, we spent on average 35 min to complete a full CEESAT assessment. Moreover, when a decision maker is faced with multiple relevant reviews, CEESAT could be used to identify the most methodologically robust reviews. Valuable lessons can also be drawn from CEESAT to improve the overall reliability of future restoration reviews. For example, we detected that some of the greatest weakness in review methods for both narrative syntheses and meta-analyses we evaluated concerned searching in multiple bibliographic sources and making the including-studies stage repeatable and transparent. These and other critical aspects that lack current, high-quality evidence syntheses can be identified with CEESAT, potentially directing new research and guiding decision makers trying to understand where the weaknesses in reviews are likely to be.

Despite the many advantages of following SR guidelines (such as CEESAT) to conduct reliable reviews, this approach is admittedly time- and resource-intensive because of the requirement to coordinate multiple reviewers, process large numbers of retrieved documents, and involve a team of expert advisors (Haddaway et al. 2015; CEE 2018). Indeed, a typical SR can take several months and may require a considerable investment (McGowan & Sampson 2005; CEE 2013). Such resource demands can make following full SR standards prohibitive for some researchers or organizations operating on limited budgets or tight time frames. Therefore, when the researcher simply lacks the resources to conduct a formal SR, it is still worthwhile to adopt many of its principles to improve the reliability of a traditional literature review (Haddaway et al. 2015; Berger-Tal et al. 2018). There may also be situations in which traditional reviews can be appropriate if a review need not be exhaustive (e.g., in exploratory and configurative reviews [Gough et al. 2012]). There are even situations in which less rigorous reviews (reaching relatively average-good CEESAT scores) could still inform decisions if they are treated with appropriate caution.

Limitations

Our evaluations (Table 3) impart confidence in accessing review reliability through CEESAT, suggesting that other restoration practitioners could apply this scoring tool themselves when selecting rigorous reviews. Nonetheless, CEESAT must be interpreted as a relatively crude measure because it does not consider some key aspects of reliability, such as analysis of analytical errors or misconduct or interpretation errors (Woodcock et al. 2014). Therefore, we echo recommendations by others

(O'Leary et al. 2016), that users refer to Woodcock et al. (2014) for a detailed consideration of using CEESAT scoring criteria.

We emphasize that only 2 meta-analyses and 3 narrative syntheses in the reviews we examined applied the full CEE SR standards (McDonald et al. 2010; Bernes et al. 2015; Gutiérrez Rodriguez et al. 2016; Land et al. 2016; Eales et al. 2018). Consequently, only these syntheses were expressly intended to meet the optimal CEESAT scores.

Toward Reliable Reviews in Restoration Ecology

Undoubtedly, evidence reviews have a crucial role to play in guiding restoration and conservation actions (Sutherland et al. 2004; Pullin & Knight 2009), but restoration syntheses vary widely in their methodological rigor. We stressed the importance of transparency, comprehensiveness, and repeatability in ensuring that reviews are reliable by discussing how the evidence-review methods used in restoration ecology have been performing relative to CEESAT criteria.

Accordingly, authors could increase the reliability of their syntheses through a priori development of a review protocol and application of reporting standards, such as the ROSES form (Haddaway et al. 2018); through use of the considerable range of bibliographic sources and gray literature (failing to make use of the full body of scientific knowledge can compromise the ability to effectively reach the desired outcomes); by ensuring repeatability and transparency in all steps of the review process (allowing an independent researcher to access the robustness of all methods employed); by documenting review activities objectively to facilitate all interpretations; and by testing results for possible publication bias.

Finally, we emphasize that review authors should follow the guidelines from the Collaboration for Environmental Evidence (CEE 2018) to elucidate their understanding of review concepts and provide direction in conducting reliable reviews. Even when researchers are not conducting a full SR or meta-analysis, they can increase the reliability of their conventional reviews by applying lessons from these guidelines (Haddaway et al. 2015). These represent positive steps toward making evidence-based restoration a reliable reality.

Acknowledgments

We are grateful to the Research Foundation of São Paulo (Fapesp, grants 2013/50718-5, 2018/18416-2, and 2019/08533-4) and Fondecyt (proyect 11191021) for providing financial support. Thanks to A. Pullin for his comments on an early version of this manuscript. We also thank 2 anonymous reviewers and the editor for helpful comments.

Supporting Information

Additional information is available online in the Supporting Information section at the end of the online article. The authors are solely responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Alpert JS. 2007. Peer review: the best of the blemished. American Journal of Medicine **120:**287–288.

Aradottir AL, Hagen D. 2013. Ecological restoration: approaches and impacts on vegetation, soils and society. Advances in Agronomy **120:**173–222.

Bayliss HR, Beyer FR. 2015. Information retrieval for ecological syntheses. Research Synthesis Methods **6:**136–148.

Berger-Tal O, et al. 2018. Systematic reviews and maps as tools for applying behavioral ecology to management and policy. Behavioral Ecology **30:**1–8.

Bernes C, Jonsson B, Junninen K, Lõhmus A, Macdonald E, Müller J, Sandström J. 2015. What is the impact of active management on biodiversity in boreal and temperate forests set aside for conservation or restoration? A systematic map. Environmental Evidence **4**. https://doi.org/10.1186/s13750-015-0050-7.

CEE (Collaboration for Environmental Evidence). 2013. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. CEE, Conwy, United Kingdom. Available from http://www.environmentalevidence.org/wp-content/uploads/2018/02/Reviewguidelines-version-4.2-final-update-1.pdf (accessed June 2020).

CEE (Collaboration for Environmental Evidence). 2018. Guidelines and standards for evidence synthesis in environmental management. Version 5.0. CEE, Conwy, United Kingdom. Available from www.environmentalevidence.org/information-for-authors (accessed March 2020).

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

Cooke SJ, et al. 2018. Evidence-based restoration in the Anthropocene-from acting with purpose to acting for impact. Restoration Ecology **26:**201–205.

Côté I, Reynolds J. 2012. Meta-analysis at the intersection of evolutionary ecology and conservation. Evolutionary Ecology **26:**1237–1252.

Dave R, et al. 2019. Second Bonn Challenge progress report: application of the Barometer in 2018. International Union for Conservation of Nature, Gland, Switzerland.

Diefenderfer HL, et al. 2016. Evidence-based evaluation of the cumulative effects of ecosystem restoration. Ecosphere **7:**e01242.

Eales J, Haddaway N, Bernes C, Cooke S, Jonsson B, Kouki J, Petrokofsky G, Taylor J. 2018. What is the effect of prescribed burning in temperate and boreal forest on biodiversity, beyond pyrophilous and saproxylic species? A systematic review. Environmental Evidence **7**. https://doi.org/10.1186/s13750-018-0131-5.

Englund G, Sarnelle O, Cooper SD. 1999. The importance of data-selection criteria: meta-analyses of stream predation experiments. Ecology **80:**1132–1141.

Gates S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. Journal of Animal Ecology **71:**547–557.

Gough D, Oliver S, Thomas J. 2012. An introduction to systematic reviews. Sage Publications, London.

Grames E, Elphick C. 2020. Use of study design principles would increase the reproducibility of reviews in conservation biology.

Biological Conservation **241**. https://doi.org/10.1016/j.biocon.2019.108385.

Guan Y, Kang R, Liu J. 2018. Evolution of the field of ecological restoration over the last three decades: a bibliometric analysis. Restoration Ecology **27:**647–660.

Gurevitch J, Koricheva J, Nakagawa S, Stewart G. 2018. Meta-analysis and the science of research synthesis. Nature **555:**175–182.

Gutiérrez Rodríguez L, Hogarth N, Zhou W, Xie C, Zhang K, Putzel L. 2016. China's conversion of cropland to forest program: a systematic review of the environmental and socioeconomic effects. Environmental Evidence **5**. https://doi.org/10.1186/s13750-016-0071-x.

Haddaway N, Woodcock P, Macura B, Collins A. 2015. Making literature reviews more reliable through application of lessons from systematic reviews. Conservation Biology **29:**1596–1605.

Haddaway NR. 2017. Response to collading science-based evidence to inform public opinion on the environmental effects of marine drilling platforms in the Mediterranean sea. Journal of Environmental Management **203:**612–614.

Haddaway NR, Macura B, Whaley P, Pullin AS. 2018. ROSES reporting standards for systematic evidence syntheses: pro forma, flow diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environmental Evidence **7**. https://doi.org/10.1186/s13750-018-0121-7.

Halme P, Toivanen T, Honkanen M, Kotiaho JS, M¨onkk¨onen M, Timonen J. 2010. Flawed meta-analysis of biodiversity effects of forest management. Conservation Biology **24:**1154–1156.

Harrison F. 2011. Getting started with meta-analysis. Methods in Ecology and Evolution **2:**1–10.

Higgins JPT, Green S. 2011. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [updated March 2011] Cochrane, London. Available from https://www.handbook.cochrane.org (accessed June 2020).

Holl KD. 2017. Restoring tropical forests from the bottom up. Science **355:**455–456.

Jennions MD, Lortie CJ, Rosenberg MS, Rothstein HR. 2013. Publication and related biases. Pages 207–236 in Koricheva J, Gurevitch J, Mengersen K, editors. Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, New Jersey.

Jennions MD, Møller AP. 2002. Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. Biological Reviews of the Cambridge Philosophical Society **77:**211–222.

Koricheva J, Gurevitch J. 2014. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology **102:**828–844.

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

Lajeunesse MJ, Forbes MR. 2003. Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecology Letters **6:**448–454.

Land M, Graneli W, Grimvall A, Hoffmann CC, Mitsch WJ, Tonderski KS, Verhoeven JTA. 2016. How effective are created or restored freshwater wetlands for nitrogen and phosphorus removal? A systematic review. Environmental Evidence **5**. https://doi.org/10.1186/s13750-016-0060-0.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. Biometrics **33:**159–174.

Leimu R, Koricheva J. 2005. What determines the citation frequency of ecological papers? Trends in Ecology & Evolution **20:**28–32.

McDonald MA, McLaren KP, Newton AC. 2010. What are the mechanisms of regeneration post-disturbance in tropical dry forest? Environmental Evidence: CEE review 07–013 (SR37).

McEuen A, Styles M. 2019. Is gardening a useful metaphor for conservation and restoration? History and controversy. Restoration Ecology **27:**1194–1198.

McGowan J, Sampson M. 2005. Systematic reviews need systematic searchers. Journal of the Medical Library Association **93:**74–80.

McMahen K, Bommel J. 2020. Towards an integrated perspective of biological conservation and ecological restoration. Restoration Ecology **28:**494–497.

Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of Internal Medicine **6:**e1000097.

Okubo Y. 1997. Bibliometric indicators and analysis of research systems: methods and examples. General technical report N1997/01. Organisation for Economic Co-operation and Development, Paris.

O'Leary B, et al. 2016. The reliability of evidence review methodology in environmental science and conservation. Environmental Science & Policy **64:**75–82.

Parker TH, et al. 2016. Transparency in ecology and evolution: real problems, real solutions. Trends in Ecology & Evolution **31:**711–719.

Philibert A, Loyce C, Makowski D. 2012. Assessment of the quality of metaanalysis in agronomy. Agriculture, Ecosystems & Environment **148:**72–82.

Possingham H, Bode M, Klein C. 2015. Optimal conservation outcomes require both restoration and protection. PLOS Biology **13:**e1002052.

Pullin AS, Knight TM. 2009. Doing more good than harm – building an evidence base for conservation and environmental management. Biological Conservation **142:**931–934.

Pullin AS, Stewart GB. 2006. Guidelines for systematic review in conservation and environmental management. Conservation Biology **20:**1647–1656.

R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Reid J, Fagan M, Zahawi R. 2018. Positive site selection bias in meta-analyses comparing natural regeneration to active forest restoration. Science Advances **4:**eaas9143.

Roberts PD, Stewart GB, Pullin AS. 2006. Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. Biological Conservation **132:**409–423.

Romanelli J, Fujimoto J, Ferreira M, Milanez D. 2018. Assessing ecological restoration as a research topic using bibliometric indicators. Ecological Engineering **120:**311–320.

Romijn E, Coppus R, De Sy V, Herold M, Roman-Cuesta R, Verchot L. 2019. Land restoration in Latin America and the Caribbean: an overview of recent, ongoing and planned restoration initiatives and their potential for climate change mitigation. Forests **10**. https://doi.org/10.3390/f10060510.

Seglen PO. 1992. The skewness of science. Journal of the American Society for Information Science **43:**628–638.

Smith R. 2006. Peer review: a flawed process at the heart of science and journals. Journal of the Royal Society of Medicine **99:**178–182.

Staples TL, Dwyer JM, Wainwright CE, Mayfield MM. 2019. Applied ecological research is on the rise but connectivity barriers persist between four major subfields. Journal of Applied Ecology **56:**1492–1498.

Sutherland WJ, Pullin AS, Dolman PM, Knight TM. 2004. The need for evidence-based conservation. Trends in Ecology & Evolution **19:**305–308.

van Eck N, Waltman L. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84:**523–538.

Vetter D, Rücker G, Storch I. 2013. Meta-analysis: a need for well-defined usage in ecology and conservation biology. Ecosphere **4:**1–24.

Woodcock P, Pullin A, Kaiser M. 2014. Evaluating and improving the reliability of evidence syntheses in conservation and environmental science: a methodology. Biological Conservation **176:**54–62.

WRI (World Resources Institute). 2018. Initiative 20 X 20. WRI, Washington, D.C. Available from https://www.wri.org/our-work/project/initiative-20x20 (acessed March 2020).