Improving quantitative synthesis to achieve generality in ecology

Article

Accepted Version

Improving quantitative synthesis to achieve generality in ecology

Rebecca Spake, Rose E. O'Dea, Shinichi Nakagawa, C. Patrick Doncaster, Masahiro Ryo, Corey T. Callaghan, James M. Bullock

Abstract

Synthesis of primary ecological data is often assumed to achieve a notion of 'generality', through the quantification of overall effect sizes and consistency among studies, and has become a dominant research approach in ecology. Unfortunately, ecologists rarely define either the generality of their findings, their estimand (the target of estimation) or population of interest. Given that generality is fundamental to science, and the urgent need for scientific understanding to curb global-scale ecological breakdown, loose usage of the term 'generality' is problematic. In other disciplines, generality is defined as comprising both generalisability: extending an inference about an estimand from the sample to the population, and transferability: the validity of estimand predictions in a different sampling unit or population. We review current practice in ecological synthesis, and demonstrate that by failing to define the assumptions underpinning generalisations and transfers of effect sizes, generality often misses its target. We provide guidance for communicating nuanced inferences, and maximising the impact of syntheses both within and beyond academia. We propose pathways to generality applicable to ecological syntheses, including the development of quantitative and qualitative criteria with which to license the transfer of estimands from both primary and synthetic studies.

Keywords: applicability; external validity; meta-analysis; meta-science; relevance

MAIN

Ecologists often seek to extend inferences from their studied systems to predict phenomena in different taxonomic, spatial or temporal settings[1]. Indeed, around 40% of the top ecology journals demand that submissions are relevant for other species, ecosystems, biomes, or time periods (Appendix S1). In principle, this is a fair request, to prevent the literature from becoming a descriptive 'stamp collection' of case studies[2], with inferences limited to the sampled population. Ecologists have pursued many roads to generalities[3,4], including developing mathematical models to predict key population parameters[5,6], unifying conceptual frameworks to predict the importance of different mechanisms in different contexts[4,7], and coordinating globally distributed experiments to predict responses of ecological systems to perturbations[8]. A further road that has gained prominence in ecology over the past 30 years is the use of 'quantitative synthesis' to identify generalities about the strength and direction of ecological effects[9].

Quantitative syntheses identify, appraise, and combine data from individual studies or sites that have measured an effect of interest, typically via meta-analysis or multilevel modelling[10–14] (Box 1). Syntheses have been used to answer both basic and applied ecological questions by quantifying, for example, the effects of major environmental drivers such as climate change on ecological communities, the effectiveness of conservation actions, and evaluating the evidence for ecological and evolutionary theories[11]. Central to quantitative synthesis is the 'effect size' estimated for each study, representing the direction and/or magnitude of an effect, commonly measured using differences between categorical group means, or the strengths of association between variables. In the absence of theoretical models or distributed experiments, effect sizes enable scientists to combine, compare and organise extensive literatures using a common measurement scale, to identify generalities across taxonomic, spatial or temporal contexts[11].

Evidence from rigorous quantitative syntheses is considered to represent one of the most methodologically robust sources for testing key ecological hypotheses, disproving or corroborating theories, and informing environmental decision making[9–11]. Concurrently, an insidious myth persists that the very act of quantitatively synthesising data from diverse studies is enough to warrant claims of generality about effect sizes[11–13]. Syntheses continue to proliferate in ecology[10], and are often associated with high-impact journals and media attention, for the apparently regional or global reach of their inferences[14]. At the same time,

however, they can stimulate much scientific debate on biases and interpretation[14]. Here we argue that current approaches to quantitative synthesis often fail to make valid inferences about the generality of effect sizes, or allow such inferences to be drawn by readers.

ASSESSING GENERALITY

Any assessment of generality requires two decisions: i) what type of generality we wish to pursue, defined by the particular target context (the population or unit of observation of interest), and ii) the estimand of interest: the quantity we have estimated from a sample, based on our research question, and that we wish to predict in the target context[15,16]. Human behavioural and health disciplines tend to define generality, or more formally, 'external validity' as the extent to which estimands drawn from a studied sample can be used to predict the same estimands of a broader population or other target contexts. The estimand might be a descriptive sample statistic of a variable of interest (e.g. mean, variance of species richness), or a measure denoting the magnitude and/or direction of a particular effect (e.g., difference in mean species richness of logged and unlogged forest stands) for a specified individual unit or population. We focus on the latter in this review.

In contrast, generality is rarely defined in ecology, with researchers often discussing the degree to which study 'findings', or 'results' can be 'transferred', 'extrapolated', 'generalised', 'applied' or 'are relevant' to other contexts. Figure 1 summarises two types of generality: generalisability and transferability[13,17,18]. Generalisability concerns the validity of extending an inference about an estimand from the sample to the sampled population. For example, ecologists might reasonably conclude that the mean effect of forest logging on understory vegetation observed in a randomly selected sample of independent forest stands in a national park in central Japan represents the mean effect across all forest stands in the park. Extending inferences beyond the sampled population extends the scope of statistical inference to different sampling units or a spatiotemporally different population of units. The validity of this extension is termed 'transferability'[18,19]. For example, one might predict a similar effect of logging to that observed in central Japan for a similar forest type in the UK. The validity or bias of this transfer could be defined as the accuracy of an estimand in a target context, quantified by the difference between the transferred estimand and the 'true' estimand.

Ecologists' statements concerning generality in both primary case studies and syntheses often do not use formal definitions of generality, and, in our experience, usually gloss over the assessments required to individuate both the studied context and the target context over which to transfer specific estimands of interest. In quantitative synthesis, the estimand is the target of estimation by an effect-size metric. A recurrent criticism is that combining effect sizes from very different contexts ('mixing apples with oranges') makes for questionable interpretability of overall ecological effects[20], leaving us with precise answers to vague questions[21]. We argue that the direction of progress in synthesis science needs resetting to enable the valid transfer of estimands in ecology. Determining the criteria or conditions that permit transfer to a specific target context is a research agenda in its own right. In this Perspective, we first examine current practice of quantitative synthesis, to understand whether and how it can substantiate claims about the generalisability or transferability of ecological effect sizes. We then provide guidance to enable nuanced inferences about the generalisability and transferability of estimands. While our focus is on synthetic research, the ability of syntheses to make general claims will depend on generality being precisely defined within primary studies, and therefore our recommendations extend to primary studies too. Finally, we outline a research agenda to guide both fundamental and applied ecological research towards valid generalisations and transfers.

Figure 1. Generality - which we use synonymously with external validity - comprises the generalisability and transferability of estimands drawn from primary and synthetic research. Syntheses collate data from primary studies, each of which usually has a well-defined and narrow context relative to the context of the synthesis, and these studies are here each represented by a fruit of one of several types. Collated, these studies form a reference sample from a hypothetical population of studies, which together cover a broader context (here of fruits, either implicitly or explicitly defined by the researcher). Generalisability concerns the validity of an inference based on a sample that is randomly or non-randomly drawn from the target population (left column). Transferability concerns the validity of inferences based on a reference sample, when applied to either a different target population or unit (target context). Transfer across space is shown as an example, to sites in a different spatial location (middle row), or an individual target site from a different population (bottom row), which may also differ in temporal or taxonomic context to the reference sample. In both cases, the synthesised samples and the populations may have well-defined or poorly defined contexts. Here, the context of the synthesis is represented by the distribution of individual studies (fruits) within three measured or unmeasured dimensions of parameter space, e.g. edaphic, taxonomic, climatic variables (*V*) that vary depending on context and may influence the outcome of a study. In our example, the hypothetical reference and target contexts overlap (within the parameter space shaded blue) despite being on different continents.

DO CURRENT PRACTICES IN QUANTITATIVE SYNTHESIS SUPPORT GENERALITY?

Quantitative syntheses, whether by meta-analysis or full-data analysis (Box 1; Table 1), generally involve some or all of three steps: i) the estimation of study-level effect sizes and an overall mean effect size, ii) estimation of heterogeneity statistics that describe differences in study-level effect sizes, and iii) attribution of effect-size heterogeneity to meaningful predictors (known as moderators), intended to provide a more nuanced configurative account of the overall effect. In syntheses, the estimand of interest is the effect size. Here we review these steps to demonstrate how current practices often do not support valid inferences about the generalisability and transferability of effect sizes.

Step 1: Estimating mean effects across a sample of primary studies

Meta-analyses of primary studies typically synthesise study-level differences between categorical treatments (e.g. Hedges' g and log response ratios LR), or the magnitudes of these changes against a continuous predictor (e.g. Pearson's z), whereas full-data analyses are performed with raw, site-level observations using (generalised) linear mixed models. Standard statistical procedures are used to estimate a measure of central tendency in effect sizes, which correspond to a weighted mean effect (meta-analysis) or a fixed effect estimated by the partial pooling of random slopes (full-data analysis). Weighting and shrinkage increase the precision of model parameters for meta-analysis and full-data analysis, respectively[22,23].

Implicitly or explicitly, these mean-effect-size estimates are generalised by the researcher from the sample of primary studies to some hypothetical population of studies, which is rarely defined. In the absence of its characterisation, it is typically implied or assumed that the target population is either i) exactly the study sample (in which case generalisation is unnecessary), or ii) the whole population from which the study observations have been randomly and independently sampled (in which case generalisation is valid). In both cases, it is assumed that the target population is implicitly defined by the inclusion and exclusion criteria of the study[19]. The validity of generalisation depends on representativeness (increased by unbiased random sampling) and sample size. Often syntheses claim to be 'global' (Figure S1), implying that inference can be generalised to some global population of studies. Such inferences are criticised when study contexts do not comprise a random and representative sample of possible contexts across a hypothetically 'global' population, due to taxonomic and

geographic biases[24]. Samples are further distorted by language[25] and publication[26] biases (e.g. file-drawer effects[27]). Moreover, mean estimates can be strongly skewed by outlying effects[28].

With at least a qualitative evaluation of possible sources of bias, such syntheses nevertheless have value. Indeed, as Rothman et al.[29] argue, "It is not representativeness of the study subjects that enhances the generalisation, it is knowledge of specific conditions and an understanding of mechanism that makes for a proper generalisation." Accordingly, the main issue is failure to characterise the reference or target contexts, even if they are narrow in scope (e.g. a limited geographic area or number of taxonomic groups studied). Rather than representativeness, a greater cause for concern is the biases introduced through the uncritical application of synthesis methods, originally developed for orthogonal medical and social studies[30,31]. For example, in serving to increase the precision of estimated mean effects, the weighting and shrinkage imposed by under-parameterised meta-analytic and multi-level models can amplify any within-study biases[30]. This is due to non-random variation in scale across studies, yielding precise yet inaccurate effect-size estimates[30]. Ecological studies employ a range of study and analytical designs[30,32]; variously factoring confounding variability in or out. A meta-analyst typically equates the different covariate configurations and study designs of primary studies when estimating effect sizes from treatment group means, and so introduces differing degrees of omitted variable bias and internal validity among the included primary studies.

Step 2: Estimating heterogeneity

The mean effects reported by a synthesis cannot be properly interpreted without an analysis of heterogeneity, or inconsistency, among effect sizes[33]. For meta-analysis, the I^2 statistic represents the percentage of variance between effect sizes that cannot be attributed to sampling error[34]. For full-data analyses, heterogeneity can be assessed using measures of random-slope variance[36,37]. Reviews have found that a large proportion of meta-analyses in ecology and evolution do not report heterogeneity statistics[35,36], and/or present aggregated mean effects that can conceal variability even within relatively homogeneous subgroups[40]. Yet heterogeneity is critical to interpreting mean effects[34]. For example, consider that a mean effect of zero biodiversity change with land use change can be achieved under two circumstances: i) effect sizes are all zero (homogenous; low between-study variance), or ii) effect sizes are very different but centred on zero (heterogeneous; high between study variance), with high heterogeneity signalling a need to explore the nature or drivers of the

variation. It is important to present the range and variability of effect sizes alongside main effect interpretation, using e,g., orchard plots[37] (e.g. as in refs 38,39), and density plots (as in refs 40, 41).

Ecological syntheses that estimate between-study variability often report very high heterogeneity (I^2 values ~90%)[42], and random slope variances[43]. Average effect sizes with high heterogeneity have questionable meaning. While meta-analysis of a set of similar experiments on a single species has a clear interpretation, interpreting a meta-effect across species and biogeographic contexts may be questionable[44]. Even Glass, an early proponent of meta-analysis[45], suggested that while meta-analysis is able to provide a "big fact", it cannot give more "sophisticated answers; they aren't there" [46]. The key point here is that while average effects are often assumed to yield generalities, averages of highly heterogenous effect sizes are neither generalisable nor transferable by themselves.

Step 3: Attributing variation to meaningful predictors

The next, and arguably the most useful, step is to attribute effect-size variation to meaningful predictors, and reach beyond the scope of individual studies to evaluate what Cooper[47] called "review-generated evidence". In meta-analysis, this is achieved by subgroup analyses that estimate and compare mean effects across meaningful groupings of studies, and the meta-regression of effect sizes against 'effect modifiers', or 'moderators'. In full-data analyses, attribution is either done by fitting more complex models that contain interaction terms between study-level or site-level covariates (e.g. that comprise an environmental gradient), or post-hoc, through regressions of random slopes on effect modifiers[48].

Attribution attempts to make inferences about the degree of transferability of an effect size, with moderators specifying the conditions to which effects can be transferred. No single reference study or sample of studies will transfer perfectly to another target context, due to inherent contextual and study-design differences. Attribution should force us to define the populations to which we wish to transfer our effect sizes (subgroups of studies, levels of predictors in a meta-regression). Target contexts are typically coarsely parameterised, however, and researchers usually estimate overall effects across broad and heterogeneous subgroupings. Obviously, subgrouping and model complexity are limited by sample size, and data availability/reporting by primary studies[49]. Attribution is prone to bias and spurious effect modification when there is covariation amongst study-design attributes (e.g. replication), random effects and effect modifiers[30]. Because the effect modifiers that

implicitly represent target contexts are often poorly characterised or heterogeneous, this limits the transferability of meta-estimates to any single setting[11].

PATHWAYS TO GENERALITY WITH ECOLOGICAL SYNTHESIS

We have demonstrated that ecology currently lacks frameworks with which to generalise or transfer estimands from quantitative syntheses. Generalisation is rarely achievable given that samples are typically non-random and heterogeneous in ecology[50]. In this section, we propose three actions that can be taken immediately by ecologists to facilitate greater nuance in communicating the transferability of the estimands. We then detail four urgent research agendas required to improve the validity of estimand transfers.

THREE ACTIONS FOR COMMUNICATING THE TRANSFERABILITY OF ESTIMANDS

Define the estimands and target contexts in a 'Constraints on Generality' statement

Psychology researchers have called for journals to require 'Constraints on Generality' (CoG) statements in the discussion sections of empirical articles, encouraging authors to draw conservative inferences, rather than make broad generalisations about undefined or ill-defined target contexts. CoG statements describe and justify target contexts, and specify assumptions the authors consider necessary for the estimand to validly transfer to other contexts[51,52]. They discourage exaggerated generality claims. CoG statements function to help both researchers and readers transfer estimands to specific target contexts. We provide an example in Box 2.

A CoG statement can explicitly define the estimand to be transferred, the target context, and any boundary conditions to which findings can be confidently applied, distinguishing between so-called 'known' and 'speculative' inferences[51]. Context parameterization might be quantitative (e.g. stating climatic, edaphic and topographic ranges), or qualitative (e.g. insects in coniferous forests of central Japan, but not all animals over the globe). Variables include those that might alter the importance of a mechanism through which a causal effect operates[53]. Context parametrisation permits both researchers and readers to implement what social scientists term the 'proximal similarity model' *sensu* Campbell 1986[54]. This model involves conceptualisation of potential target contexts as a gradient of similarity, from most closely similar to least similar. Proximal similarity supports transferability to those

populations that are spatially, temporally, and taxonomically most alike (i.e., most proximally similar to) those in the focal study[13].

Researchers could make statements about the predicted estimand in a specific target context, e.g. the magnitude and sign of an effect on a specified outcome, and how estimands might change along a given gradient under specified conditions, and state whether the target gradient extends beyond the range of the reference population's parameter space. Researchers could articulate assumptions underlying the predictions (e.g. what conditions must hold, such as site historical factors), as well as potential ecological and/or societal impacts of an assumption being violated.

We see an opportunity for reviewers to be involved in improving CoG statements. If the onus is only on authors to specify generality, these statements risk being arbitrarily subjective, and marginalised to a perfunctory 'limitations' section. Reviewers could serve two roles in this regard. First, at the stage of submitting their evaluation, reviewers could be asked a short-response question about what they perceive the generalisability and transferability of the empirical findings to be. If the statements of the authors and reviewers diverge notably, this would indicate to the editor a lack of clarity in the manuscript about generality or necessary context. For journals that provide peer-review reports alongside published papers, the reviewers' perceptions of generality could provide additional insights to readers. Second, reviewers can serve a role, again through a short-response question, in discouraging authors from exaggerating generality, especially in the title and abstract.

Move beyond static representations of ecological relationships

Researchers could work harder to meaningfully communicate contingency, uncertainty and transferability of estimands to different audiences, including researchers and practitioners. In both primary and synthetic studies, the usual current practice is to display outputs of analyses as two-dimensional (2D) static plots, typically holding other covariates at their mean values[55]. Given the conditional character of ecological relationships, estimated using nonlinear link functions and linear models with interaction terms, such 2D plots are often ineffective at displaying the range and variability of estimands[56]. Possible alternatives include interactive graphics that enable readers to explore underlying data points from full-data syntheses, and the prediction of marginal effects for user-specified covariate values (e.g. ref 57). For example, McCabe et al.[58] produced an interactive web application to help psychology researchers visualise interaction effects, and communicate the statistical integrity of analyses

(https://connorjmccabe.shinyapps.io/interactive/). For meta-analysis, 'dynamic meta-analysis' software has been developed, whereby effect sizes can be filtered and weighted, and results can be recalculated, using subgroup analysis, meta-regression, and recalibration[59], which could be extended to alternate weighting schemes that incorporate generality criteria[31,60]. EviAtlas is an example of open source software for producing interactive visualisations of systematic map databases[61]. These applications could be embedded within online publications, which increasingly support interactive graphics and code[62,63].

Quantify the 'transfer domain' for full-data syntheses

In addition to quantitative context parameterisation (Action 1), researchers could identify the 'transfer domain' that delineates the parameter space to which effect sizes can be validly transferred (given CoG statements and assumptions), also known as the 'applicability domain', in predictive modelling across disciplines including chemistry[64], material science[65], and environmental science[66]. For full-data syntheses of large datasets, cross-validation techniques could be used, wherein model parameters are estimated using 90% of the primary studies (training set), and model predictive performance evaluated using the remaining 10% (test set). After repeating on different combinations of primary studies in training and test sets, studies for which effect sizes are not predicted well would be considered outside of the transfer domain. To identify the boundary conditions, one could identify the characteristics of studies that are unpredictable. Employing cross validation for meta-analysis will change the focus from the most precise estimate and its statistical significance, to how well estimands transfer to different contexts.

FOUR AGENDAS FOR DEVELOPING A SCIENCE OF GENERALITY APPLICABLE TO SYNTHESIS

Here we propose four research agendas to guide the development of both quantitative and qualitative assumptions that underpin the generalisability and transferability of estimands for scientists and policymakers. We identify six key steps (Figure 2) which could help to formalise the assumptions that underpin transfer of estimands to specific contexts in ecology[67].

Figure 2. Six steps to transfer an estimand to a target context

Develop qualitative and quantitative criteria with which to evaluate transferability of an estimand (for scientists)

After specifying an estimand for a target context of interest (Figure 2, steps 1 and 2), researchers could develop qualitative criteria or quantitative indicators with which to appraise the transferability, or assumptions that (if met) justify the transfer of an estimand of interest (step 3). These criteria can be used to enhance CoG statements (Action 1) and guide the appraisal of primary studies that are used in quantitative syntheses. Criteria could comprise descriptors of dissimilarity between the reference and target contexts (their covariate distributions), study-design attributes (e.g. replication, spatial interspersion), analytical design attributes (e.g. model complexity, statistical matching), modelling choice (e.g. machine learning), and the mechanistic nature of the causal relationships. Ideally, these criteria and assumptions would be identified at the beginning of a study, to guide its design, rather than at the end[68]. While high-level categories of appraisal criteria are likely to be useful to guide the analysis and interpretation of primary and synthetic studies, exact criteria will be specific to the ecological question and estimand of interest.

Health disciplines have developed objective criteria with which to judge the external validity of primary studies for a defined target context, e.g.[69–71]. For example, the Population-Intervention-Environment-Transfer Model of Transferability helps different audiences to judge the transferability of a health intervention, according to characteristics of the studied Population (socio-demographic, attitudinal), Intervention (internal validity of study), Environment (public perception, climate) and Transfer (feasibility of intervention)[70]. These have been recently extended to syntheses[68,72]. For example, the TRANSFER approach[68] supports collaboration between researchers and stakeholders during the review process to systematically and transparently consider factors that may influence the transferability of medical systematic review findings. To support the identification of important contextual variables with which to define reference and target contexts and evaluate the validity of potential transfers, the use of 'selection diagrams' can help identify important conditioning variables and study design attributes that might influence the transferability of causal effects. Pearl and Bareinboim[67,73] proposed the use of these graphical representations of causal relationships, which formally articulate commonalities and differences in the form of unobserved factors capable of causing differences in causal effects between reference and target contexts. This approach is a useful tool for identifying important conditioning

covariates and detailing the assumptions and tests that are required to develop qualitative indicators and tests of transferability (example in Box 3).

Develop quantitative methods to transfer estimands (for scientists)

Ecologists could develop methods to transfer estimands to different target contexts, once the reference and target contexts have been parameterized. Degtiar et al. (2021)[74] reviewed the numerous quantitative approaches that have been developed in primarily health-related disciplines to: i) evaluate the validity of transferring an estimand to a specified target context, based on a set of assumptions (Figure 2, step 3) and the quantitative dissimilarity of the study and target contexts (step 4), and ii) 'external validity bias adjustment' methods to adjust an estimand for a target context (step 5). For example, Pearl and Bareinboim formalised a range of 'transport formulae' associated with selection diagrams that enable the re-calibration of average population-level effect sizes for a well-defined target context, e.g. through re-weighting observations in the reference population in proportion to distributions of conditioning covariates in the target context[18,75] (example in Box 2). The choice of method for estimand adjustment may be restricted by data availability (e.g., summary-level *vs.* individual-level data) and mechanistic understanding of the target system.

Validation of quantitative transfers (step 6), and of the methods developed to enable transfer, will only be possible with independent studies and data using cross validation (Action 3), although they are often unavailable or insufficient for a target context. Transfer methods and understanding need development as a discipline. In the meanwhile, data gaps might be filled by making use of continental-scale, fine-resolution data from environmental monitoring programmes that span multiple environmental contexts, such as The National Science Foundation's National Ecological Observatory Network (NEON, https://www.neonscience.org), and national forest inventories. In the absence of validation data for a target contexts, transferability could be estimated by contrasting predictions with existing expert knowledge, simulations, or by performing controlled, distributed experiments[8].

Conduct interdisciplinary research that seeks to understand how multiple stakeholders perceive generalisability and transferability (for scientists and practitioners)

Scientists need to communicate the transferability, contingency and uncertainty of ecological effects in a meaningful and practicable way. This requires an understanding of how perceptions of transferability and uncertainty are formed by different audiences, including

scientists, practitioners and policymakers[76,77]. Interdisciplinary research is required to understand how different attributes affect the perceived transferability of ecological effects (using e.g., surveys, workshops). These might include i) audience attributes (e.g., sector, experience), ii) study context (biogeography, climatic conditions), iii) study design attributes (e.g., design, scale, replication), and iv) presentation attributes (e.g., graphical presentation of results). Next, we can use this understanding to determine how uncertainty and contingencies are unambiguously communicated, by trialling different methods of translation, and for improving CoG statements.

Conduct adaptive research that feeds into syntheses (for scientists and science funders)

Research funding is usually based on competition between individual proposals, with an emphasis on novelty. Distributed experiments have become popular in many disciplines[78] as an approach that aims at generality by repeating an experimental design in multiple locations (e.g. Nutrient Network [NutNet][8], Marine Global Earth Observatory [MarineGEO][79], and ManyLabs in psychology[80]). In practice, such distributed experiments are poorly resourced, depending for setup and maintenance on freely-offered endeavours of dedicated researchers. Large-scale, multinational and long-term funding to institutions for collaboration could transform this approach, to sample across the range of contextual variables, as orthogonally as possible. Importantly, the results could inform extensions to these studies, or a new set of studies, in accordance with the concept of 'adaptive experimentation'[81]. This would lead to syntheses that inform transferable research designs in an iterative manner, rather than 'making do' with what has gone before. This idea replaces the current paradigm of individual-level competitiveness and novelty with institutional-level collaboration and scope for generality, and it provides a framework for individual scientists to develop their talents in collaborative teams.

Acknowledgements

We thank J. Chase for informative discussion of concepts, and three reviewers for valuable comments. RS is grateful for funding from the German Centre for Integrative Biodiversity Research – iDiv - Halle-Jena-Leipzig. JMB was funded under UKCEH National Capability project 06895. CTC was supported by a Marie Skłodowska-Curie Individual Fellowship (no. 891052).

Author information

Authors and affiliations

Name	Affiliation	Email address	ORCID
Rebecca Spake	School of Biological Sciences, University of Reading	becksspake@gmail.com	0000-0003-4671-2225
Rose E. O'Dea	School of Ecosystem and Forest Sciences, University of Melbourne, Melbourne, Australia	rose.eleanor.o.dea@gmail.com	0000-0001-8177-5075
Shinichi Nakagawa	Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney	s.nakagawa@unsw.edu.au	0000-0002-7765-5182
C. Patrick Doncaster	School of Biological Sciences, University of Southampton, SO17 1BJ, UK	cpd@soton.ac.uk	0000-0001-9406-0693
Masahiro Ryo	Leibniz Centre for Agricultural Landscape Research (ZALF) Brandenburg University of Technology Cottbus–Senftenberg	masahiro.ryo@zalf.de	0000-0002-5271-3446
Corey T. Callaghan	German Centre for Integrative Biodiversity research – iDiv - Halle-Jena-Leipzig, Puschstrasse 4, 04103, Leipzig, Germany	corey.callaghan@idiv.de	**0000-0003-0415-2709**
James M. Bullock	UK Centre for Ecology & Hydrology, Oxfordshire, OX10 8BB, UK	jmbul@ceh.ac.uk	0000-0003-0529-4020

Corresponding author

Correspondence to Rebecca Spake: R.Spake@reading.ac.uk

Author contributions

RS conceived the idea and developed a first draft with JMB. REO, SN, CPD, MR, and CTC contributed to idea development and paper writing.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Houlahan, J. E., McKinney, S. T., Anderson, T. M. & McGill, B. J. The priority of prediction in ecological understanding. *Oikos* **126**, 1–7 (2017).

2. Lawton, J. H. Are there general laws in ecology? *Oikos* **84**, 177–192 (1999).

3. Elliott-Graves, A. Generality and Causal Interdependence in Ecology. *Philos. Sci.* **85**, 1102–1114 (2018).

4. Fox, J. W. The many roads to generality in ecology. *Philos. Top.* **9**, 83–104 (2019).

5. McGill, B. J. *et al.* Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* **10**, 995–1015 (2007).

6. MacArthur, R. H., and E. O. W. An equilibrium theory of insular zoogeography. *Evolution (N. Y).* **17**, 373–87 (1963).

7. Gurevitch, J., Fox, G. A., Wardle, G. M., Inderjit & Taub, D. Emergent insights from the synthesis of conceptual frameworks for biological invasions. *Ecol. Lett.* **14**, 407–418 (2011).

8. Borer, E. T. *et al.* Finding generality in ecology: A model for globally distributed experiments. *Methods Ecol. Evol.* **5**, 65–73 (2014).

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

10. Anderson, S. C. *et al.* Trends in ecology and conservation over eight decades. *Front. Ecol. Environ.* **19**, 274–282 (2021).

11. Kneale, D., Thomas, J., O'Mara-Eves, A. & Wiggins, R. How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision making? *Res. Synth. Methods* **10**, 44–56 (2019).

12. Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R. & Dalton, C. M. Debunking myths and urban legends about meta-analysis. *Organ. Res. Methods* **14**, 306–331 (2011).

13. Polit, D. F. & Beck, C. T. Generalization in quantitative and qualitative research: Myths and strategies. *Int. J. Nurs. Stud.* **47**, 1451–1458 (2010).

14. Cardinale, B. J., Gonzalez, A., Allington, G. R. H. & Loreau, M. Is local biodiversity declining or not? A summary of the debate over analysis of species richness time trends. *Biol. Conserv.* **219**, 175–183 (2018).

15. Lundberg, I., Johnson, R. & Stewart, B. M. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *Am. Sociol. Rev.* **86**, 532–565 (2021).

16. Lawrance, R. *et al.* What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *J. Patient-Reported Outcomes* **4**, (2020).

17. Findley, M. G., Kikuta, K. & Denly, M. External validity. *Annu. Rev. ofPolitical Sci.* **24**, 365–393 (2021).

18. Pearl, J. & Bareinboim, E. External validity: From do-calculus to transportability across populations. *Stat. Sci.* **29**, 579–595 (2014).

19. Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R. & Stuart, E. A. Target Validity and the Hierarchy of Study Designs. *Am. J. Epidemiol.* **188**, 438–443 (2019).

20. Carpenter, C. J. Meta-analyzing apples and oranges: How to make applesauce instead of fruit salad. *Hum. Commun. Res.* **46**, 322–333 (2020).

21. Rohrer, J. M. & Arslan, R. C. Precise Answers to Vague Questions: Issues With Interactions. *Adv. Methods Pract. Psychol. Sci.* **4**, (2021).

22. Breslow, N. E. & Clayton, D. G. Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.* **88**, 9 (1993).

23. Koricheva, J. & Gurevitch, J. Uses and misuses of meta-analysis in plant ecology. *J. Ecol.* **102**, 828–844 (2014).

24. Gonzalez, A. *et al.* Estimating local biodiversity change: A critique of papers claiming no net loss of local diversity. *Ecology* **97**, 1949–1960 (2016).

25. Konno, K. *et al.* Ignoring non-English-language studies may bias ecological meta-analyses. *Ecol. Evol.* **10**, 6373–6384 (2020).

26. Nakagawa, S. *et al.* Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods Ecol. Evol.* **13**, 4–21 (2022).

27. Rosenthal R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979).

28. Leung, B. *et al.* Clustered versus catastrophic global vertebrate declines. *Nature* **588**, 267–271 (2020).

29. Rothman, K. J., Gallacher, J. E. J. & Hatch, E. E. Why representativeness should be avoided. *Int. J. Epidemiol.* **42**, 1012–1014 (2013).

30. Spake, R. *et al.* Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecol. Lett.* **24**, 374–390 (2021).

31. Spake, R. & Doncaster, C. P. Use of meta-analysis in forest biodiversity research: key challenges and considerations. *For. Ecol. Manage.* **400**, 429–437 (2017).

32. Christie, A. P. *et al.* Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *J. Appl. Ecol.* **56**, 2742–2754 (2019).

33. Nakagawa, S., Noble, D. W. A., Senior, A. M. & Lagisz, M. Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biol.* **15**, 1–14 (2017).

34. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).

35. Lorah, J. Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large-Scale Assessments Educ.* **6**, (2018).

36. Schielzeth, H. & Nakagawa, S. Conditional repeatability and the variance explained by reaction norm variation in random slope models. *bioRxiv* (2020) doi:10.1101/2020.03.11.987073.

37. Nakagawa, S. *et al.* The orchard plot: Cultivating a forest plot for use in ecology, evolution, and beyond. *Res. Synth. Methods* **12**, 4–12 (2021).

38. Ojha, M., Naidu, D. G. T. & Bagchi, S. Meta-analysis of induced anti-herbivore defence traits in plants from 647 manipulative experiments with natural and simulated herbivory. *J. Ecol.* 1–18 (2022) doi:10.1111/1365-2745.13841.

39. Dodds, K. C. *et al.* Material type influences the abundance but not richness of

colonising organisms on marine structures. *J. Environ. Manage.* **307**, (2022).

40. O'Connor, M. I. *et al.* A general biodiversity–function relationship is mediated by trophic level. *Oikos* **126**, 18–31 (2017).

41. Dornelas, M. *et al.* Assemblage time series reveal biodiversity change but not systematic loss. *Science (80-.).* **344**, 296–299 (2014).

42. Senior, A. M. *et al.* Heterogeneity in ecological and evolutionary meta- analyses: its magnitude and implications. *Ecology* **97**, 3293–3299 (2016).

43. Blowes, S. A. *et al.* The geography of biodiversity change in marine and terrestrial assemblages. *Science (80-.).* **366**, 339–345 (2019).

44. Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biol. Rev.* **82**, 591–605 (2007).

45. Glass, G. V. Primary, Secondary, and Meta-Analysis of Research. *Educ. Res.* **5**, 3 (1976).

46. Glass, G. V. Meta-analysis at 25. http://www.gvglass.info/papers/meta25.html (2000).

47. Cooper, H. M. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowl. Soc.* **1**, 104–126 (1988).

48. Soranno, P. A. *et al.* Cross-scale interactions: Quantifying multi-scaled cause-effect relationships in macrosystems. *Front. Ecol. Environ.* **12**, 65–73 (2014).

49. Gerstner, K. *et al.* Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods Ecol. Evol.* **8**, 777–784 (2017).

50. Hortal, J. *et al.* Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 523–549 (2015).

51. Simons, D. J., Shoda, Y. & Lindsay, D. S. Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspect. Psychol. Sci.* **12**, 1123–1128 (2017).

52. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* 1–37 (2020) doi:10.1017/S0140525X20001685.

53. Lopez, P. M., Subramanian, S. V. & Schooling, C. M. Effect measure modification

conceptualized using selection diagrams as mediation by mechanisms of varying population-level relevance. *J. Clin. Epidemiol.* **113**, 123–128 (2019).

54. Campbell, D. T. Relabeling internal and external validity for the applied social sciences. in *Advances in QuasiExperimental Design and Analysis* (ed. Trochim, W.) 67–77 (Jossey-Bass, 1986).

55. Mize, T. D. Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociol. Sci.* **6**, 81–117 (2019).

56. Karaca-Mandic, P., Norton, E. C. & Dowd, B. Interaction terms in nonlinear models. *Health Serv. Res.* **47**, 255–274 (2012).

57. Spake, R. *et al.* Forest damage by deer depends on cross-scale interactions between climate , deer density and landscape structure. 1376–1390 (2020) doi:10.1111/1365-2664.13622.

58. McCabe, C. J., Kim, D. S. & King, K. M. Improving Present Practices in the Visual Display of Interactions. *Adv. Methods Pract. Psychol. Sci.* **1**, 147–165 (2018).

59. Shackelford, G. E. *et al.* Dynamic meta-analysis: A method of using global evidence for local decision making. *bioRxiv* 1–13 (2020) doi:10.1101/2020.05.18.078840.

60. Christie, A. P. *et al.* Innovation and forward-thinking are needed to improve traditional synthesis methods: A response to Pescott and Stewart. *J. Appl. Ecol.* 1–7 (2022) doi:10.1111/1365-2664.14154.

61. Haddaway, N. R. *et al.* EviAtlas: A tool for visualising evidence synthesis databases. *Environ. Evid.* **8**, 1–10 (2019).

62. Delory, B. M., Li, M., Topp, C. N. & Lobet, G. archiDART v3.0: A new data analysis pipeline allowing the topological analysis of plant root systems. *F1000Research* **7**, 1–14 (2018).

63. Perkel, J. M. The Future of Scientific Figures. *Nature* **554**, 133–134 (2018).

64. Weaver, S. & Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph. Model.* **26**, 1315–1326 (2008).

65. Sutton, C. *et al.* Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **11**, 1–9 (2020).

66. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12**, 1620–1633 (2021).

67. Pearl, J. & Bareinboim, E. Transportability of causal and statistical relations: A formal approach. *Proc. - IEEE Int. Conf. Data Mining, ICDM* 540–547 (2011) doi:10.1109/ICDMW.2011.169.

68. Munthe-Kaas, H., Nøkleby, H. & Nguyen, L. Systematic mapping of checklists for assessing transferability. *Syst. Rev.* **8**, 1–16 (2019).

69. Dekkers, O. M., von Elm, E., Algra, A., Romijn, J. A. & Vandenbroucke, J. P. How to assess the external validity of therapeutic trials: A conceptual approach. *Int. J. Epidemiol.* **39**, 89–94 (2010).

70. Schloemer, T. & Schröder-Bäck, P. Criteria for evaluating transferability of health interventions: A systematic review and thematic synthesis. *Implement. Sci.* **13**, 1–17 (2018).

71. Fernandez-Hermida, J. R., Calafat, A., Becoña, E., Tsertsvadze, A. & Foxcroft, D. R. Assessment of generalizability, applicability and predictability (GAP) for evaluating external validity in studies of universal family-based prevention of alcohol misuse in young people: Systematic methodological review of randomized controlled trials. *Addiction* **107**, 1570–1579 (2012).

72. Avellar, S. A. *et al.* External Validity: The Next Step for Systematic Reviews? *Eval. Rev.* **41**, 283–325 (2017).

73. Bareinboim, E. & Pearl, J. A General Algorithm for Deciding Transportability of Experimental Results. *J. Causal Inference* **1**, 107–134 (2013).

74. Degtiar, I. & Rose, S. A Review of Generalizability and Transportability. 1–30 (2021).

75. Bareinboim, E. & Pearl, J. Meta-transportability of causal effects: A formal approach. *J. Mach. Learn. Res.* **31**, 135–143 (2013).

76. Jamieson, D. Scientific uncertainty: How do we know when to communicate research findings to the public? *Sci. Total Environ.* **184**, 103–107 (1996).

77. Burchett, H. E. D., Mayhew, S. H., Lavis, J. N. & Dobrow, M. J. When can research from one setting be useful in another Understanding perceptions of the applicability

and transferability of research. *Health Promot. Int.* **28**, 418–430 (2013).

78. Forscher, P. *et al.* Build up big-team science. *Nature* **601**, 505–507 (2022).

79. Whalen, M. A. *et al.* Climate drives the geography of marine consumption by changing predator communities. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28160–28166 (2020).

80. Moshontz, H. *et al.* The Psychological Science Accelerator : Advancing Psychology through a Distributed Collaborative Network An updated version of this manuscript is published online at Advances in Methods and Practices in Psychological Science (https://doi.org/10.1177/2515. *Adv. Methods Pract. Psychol. Sci.* **1**, 501–515 (2018).

81. Marschner, I. C. A general framework for the analysis of adaptive experiments. *Stat. Sci.* **36**, 465–492 (2021).

82. Mengersen, K., Gurevitch, J. & Schmid, C. H. Meta-analysis of Primary Data. in *Handbook of Meta-analysis in Ecology and Evolution* (eds. Koricheva, U., Gurevitch, J. & Mengersen, K.) (Princeton University Press, 2013).

83. Hudson, L. N. *et al.* The database of the PREDICTS (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems) project. *Ecol. Evol.* **7**, 145–188 (2017).

84. M, D., Antão, L., F, M., AE, B. & AE, M. BioTIME: A database of biodiversity time series for the Anthropocene. *Glob. Ecol. Biogeogr.* **27**, 760–786 (2018).

85. Salguero-Gómez, R. *et al.* The compadre Plant Matrix Database: An open online repository for plant demography. *J. Ecol.* **103**, 202–218 (2015).

86. Salguero-Gómez, R. *et al.* COMADRE: A global data base of animal demography. *J. Anim. Ecol.* **85**, 371–384 (2016).

87. Pastor, D. A. & Lazowski, R. A. On the Multilevel Nature of Meta-Analysis: A Tutorial, Comparison of Software Programs, and Discussion of Analytic Choices. *Multivariate Behav. Res.* **53**, 74–89 (2018).

88. Spake, R. *et al.* Meta-analysis of management effects on biodiversity in plantation and secondary forests of Japan. *Conserv. Sci. Pract.* **1**, e14 (2019).

89. Forestry Agency of Japan. Forest Ecosystem Diversity Basic Survey (in Japanese).

Available online at: https://www.rinya.maff.go.jp/j/keikaku/tayouseichousa/index.html. (2019).

90. S, I., S, I., N, M. & GP, B. Maintaining plant species composition and diversity of understory vegetation under strip-clearcutting forestry in conifer plantations in Kyushu, southern Japan. *For. Ecol. Manage.* **231**, 234–241 (2006).

91. Utsugi, E. *et al.* Hardwood recruitment into conifer plantations in Japan: Effects of thinning and distance from neighboring hardwood forests. *For. Ecol. Manage.* **237**, 15–28 (2006).

92. Kominami, Y. *et al.* Classification of bird-dispersed plants by fruiting phenology, fruit size, and growth form in a primary lucidophyllous forest: an analysis, with implications for the conservation of fruit–bird interactions. *Ornthological Sci.* **2**, 3–23 (2003).

93. Tsujino, R. & Matsui, K. Forest regeneration inhibition in a mixed broadleaf-conifer forest under sika deer pressure. *J. For. Res.* **00**, 1–6 (2021).

94. Spake, R., Soga, M., Catford, J. A. & Eigenbrod, F. Applying the stress-gradient hypothesis to curb the spread of invasive bamboo. *J. Appl. Ecol.* **58**, 1993–2003 (2021).

95. Clark, M. Shrinkage in mixed effects models. Available at: https://m-clark.github.io/posts/2019-05-14-shrinkage-in-mixed-models/. Last accessed 18 November 2020. (2019).

96. Gurevitch, J. & Hedges, L. V. Statistical issues in ecological meta-analyses. *Ecology* **80**, 1142–1149 (1999).

Box 1. Current practices in quantitative synthesis

Two approaches to quantitative synthesis are widely used: (a) the meta-analysis of study-level summary statistics (hereafter 'meta-analysis'), which requires treatment-level means, standard deviations and sample sizes; (b) full-data analyses that fit multilevel (generalised) linear mixed models to raw, site-level observations, hereafter 'full-data analysis' (Table 1). In health disciplines, full-data analyses are known as 'individual patient data meta-analysis', and are considered the 'gold standard'[82], due to their potential for resolving issues regarding study-specific designs and confounding variation. The use of full-data analyses has also surged in ecology, aided by open-science policies that encourage or mandate the publication of raw data alongside articles, and initiatives that collate raw data (e.g., PREDICTS[83], BioTime[84], COMPADRE/COMADRE[85,86]). While definitions vary within and between disciplines, e.g. meta-analysis may be considered a special case of multilevel modeling[87], we use the term 'synthesis' to encompass both meta-analysis and full-data analysis, as defined in Table A.

Table A. Two approaches to the synthesis of primary studies that have measured responses of some ecological variable *Y*, such as biodiversity or carbon storage, to variable *X*, and effect modification by variable *Z*.

	Meta-analysis	Full-data analysis
Input data	Study-level summary statistics (mean, standard deviation, *n*) compiled across multiple studies. Primary studies may have measured outcomes in different units.	Study-level raw data compiled across multiple studies. Unit of measurement must be consistent across studies.
Study-level effect sizes	study-level differences between categorical treatments (e.g. Hedges' *g* or log response ratios), or the magnitudes of these changes against a continuous predictor (*e.g.* correlations).	Study-level random slopes on the scale of the linear predictor
Statistical procedure	Precision-weighting, generally using the inverse of the sum of study-level and between-study variance.	Partial pooling, wherein group (study) estimates are 'shrunk' toward the population mean as a function of the relative variance of each estimate.
Estimate of overall mean effects	Meta-estimate of mean effect (*ΔY*; top left in figure)	Fixed-effect estimate (top right in figure)
Estimate of between-study heterogeneity	Heterogeneity statistics e.g. I^2. Benchmarks of I^2 of 25%, 50%, and 75% are interpreted as small, medium, and high, respectively.	Concurrent interpretation of three parameters: the variance of i) random slopes ii) random intercepts, and iii) the covariance of intercepts and slopes.
Attribution	Comparison of subgroup mean effects, or meta-regression of effect sizes on meaningful *'effect modifiers'* or *'moderators' (Z;* bottom left in figure).	The analyst may fit an interaction term between the *X* and *Z*, and interrogate the marginal effects. Sometimes analysts perform *post-hoc* analyses of random slopes, e.g. regression on '*effect modifiers'* or *'moderators'*, *(Z*; bottom right in figure*)*.

Box 2. Example 'Constraints on Generality' statement for synthesis of plantation thinning effects on broadleaved sapling abundance

Summary of study: Spake et al. (2019)[88] synthesised the effects of stand-level forest management interventions on biodiversity in Japan. Here we present effect sizes representing the effect of plantation thinning on broadleaved tree regeneration, for plantations dominated by either *Cryptomeria japonica* (sugi) or *Chamaecyparis obtusa* (hinoki) distributed across Japan. For each comparison, a log response ratio was estimated to represent the proportionate difference in broadleaved sapling/seedling abundance between replicates of thinned and unthinned stands. Effect sizes were meta-regressed on thinning intensity, measured as the percent of stand volume removed. A positive effect of stand thinning on biodiversity increased with thinning intensity (below, left). Further details are available in Appendix S2.

Left: Effect sizes representing the effects of plantation thinning on abundance of saplings and seedlings depend on thinning intensity, showing grey-shaded 95% CI in the regression based on between-study and within-study uncertainty; values above horizontal dotted line signify higher abundance in thinned than unthinned stands. Point colour and shape combinations correspond to study identifiers, while point size is proportional to estimated weights. Middle: Spatial distribution of study sites in Japan. Right: Distribution of studies in parameter space according to mean annual rainfall and elevation. Coloured study locations are overlain on parameter space occupied by plots dominated by sugi or hinoki surveyed in a national forest inventory[89] (grey shading corresponds to plot density; see Appendix S2 for details).

Constraints on Generality: Reductions in sugi and hinoki stand volumes by greater than 30% are likely to increase sapling and seedling abundances in young, even-aged plantations between 20 and 41 years old, located across warm-temperate Japan (above figure, middle & right). For these closed-canopy forests, the positive effect of thinning on sapling abundance should increase with thinning intensity, up to 60%. Further studies are required to establish whether positive effects remain or indeed become stronger after 60%, because planted trees might have indirect effects on broadleaved regeneration: clear-cutting (100% reductions) can lead to dominance of herbs and/or shrubs, which inhibit the regeneration of broadleaved tree species[90]. In the studies collated, stands had been surveyed between two and seven years after line or selective thinning. Positive effects may not be evident after longer periods, as recruitment to older age classes may not persist following rapid canopy closure, and repeated thinning may be required to ensure the survival of regenerated seedlings.

Positive effects of thinning on broadleaved tree regeneration should hold for plantations with intact broadleaved seed banks, which are major sources of seedlings recruited after disturbance in conifer plantations[91], and for sites located in highly-forested landscapes. We caution against transferring the positive effect of thinning to landscapes with little forest cover, because recruitment has been shown to decline with distance to forest[91], with seeds of more than 60% of tree species in warm-temperate forests of Japan dispersed by forest-dwelling birds[92]. We speculate that these positive effects will extend to closed-canopy plantations in other temperate regions where light availability is the most limiting resource for understory plants, but caution that the positive effect of thinning will likely not extend to older plantations with more complex age structures and open canopies, i.e. to stands with forest floors that are not light-limited, or to stands in regions with high densities of deer (*Cervus japonicus*) that limit

regeneration[93], or where thinning is known to enhance single-species dominance or invasive species establishment (e.g., giant bamboo [*Phyllostachys* sp.] in warm-temperate Japan)[94]

Box 3: Selection diagram approach for identifying contextual variables and assumptions, and transport formulae to enable transfer of an estimand.

Selection diagram approach for the effect of forest thinning on understory biodiversity (Adapted from Pearl & Bareinboim (2013)[75].

a) Consider the problem of transporting experimental results between two locations. We first conduct a randomized experiment in a location (reference context) and estimate the causal effect of forest thinning (treatment *X*) on understory biodiversity (outcome *Y*) for every stand age group (*Z* = *z*), denoted *P*(*y*|*do*(*x*),*z*). We now wish to transport the results to forests in another location (target context), but we find the distribution *P*(*xyz*) to be different from the one in target context (call the latter *P**(*xyz*)). In particular, the average age of the trees is significantly lower than that in the reference context. How do we estimate the causal effect of *X* on *Y* in the target context, denoted *R* = *P**(*y*|*do*(*x*),*z*)?

b) The selection diagram conveys the assumption that the only difference between the two populations are factors determining age distributions of trees shown as *S* → *Z,* while age-specific effects *P*(*y*|*do*(*x*),*Z*=*z*) are invariant across forest contexts. Dashed arcs (e.g., X⇠⇢Y) represent the presence of latent variables affecting both *X* and *Y*.

Under these assumptions, the causal effect in the target context, *R,* can be estimated using a transport formula as follows:

$R = \sum_Z P^*(y|do(x),z)\,P^*(z)$

$= \sum_Z P(y|do(x),z)\,P^*(z)$

It combines experimental results obtained in the reference context, *P*(*y*|*do*(*x*),*z*); with observational aspects of target context *P*z,* to obtain an experimental claim *P**(*y*|*do*(*x*)*,* about the target context. By formalising this graphically and formulaically, we are forced to define what we must assume about other confounding variables beside stand age, both latent and observed,for our formulae to have validity.

Box 4. Glossary of terms

Accuracy / bias: the distance of an estimate from the value it is estimating, with a large distance signifying low accuracy / high bias.

Boundary conditions: the regions of the parameter space that describe a context, within which an inference is valid.

Causal inference: an evidence-based conclusion about the causal, driving effect of a particular phenomenon.

Effect modification: an effect magnitude and/or direction that varies with the values of another effect, and vice versa.

Estimand: the target of estimation, characterised by: a response variable of interest (e.g. species richness), an independent variable of interest (e.g. forest logging), a summary measure (e.g., the standardised mean difference in species richness between the populations of logged and unlogged stands: $[\mu_1 - \mu_2]/\sigma$), the target population or unit of interest (e.g., planted forest stands within a national park)..

External validity: Here referred to as 'generality'. The capacity for a sample estimand to apply to a specified target population. Two types are distinguished: generalisability and transferability.

Generalisability: concerns the validity of extending an inference about an estimand from the sample to the population from which it is drawn. Generalisability could be defined as the accuracy of a sample estimand, in terms of its difference from the true population estimand.

Internal validity:. The degree to which observed covariation between a dependent and an independent variable can be interpreted as a causal effect.

Precision: the distribution of replicate estimates around their mean, with a tight distribution signifying high precision. In the absence of systematic bias, greater precision leads to higher accuracy.

Primary study: a study that gathers new data on a particular population (distinguished from a secondary study, such as a synthesis of primary studies).

Sampled population: the set of observational units of a distributed variable that define the scope of inference of the testable hypotheses. Statistical analyses require random and independent sampling from the population of interest, which means that the population needs defining at the design stage. The outcome of statistical testing (e.g., detection of a trend) applies to the sampled population, not to the sample(s). Thus, confidence intervals around a sample mean describe the range of plausible values of the population mean given the sample.

Shrinkage: a fundamental property of multilevel models, also known as 'borrowing strength', wherein individual group (e.g., study-level) estimates are shrunk toward the overall population mean. Data nuances will determine the relative amount of strength borrowed per study, but in general, shrinkage is a function of the relative variance of each estimate, and is greater for groups with extreme values and lower replication[95]. As with weighting in meta-analyses of effect sizes, shrinkage functions to reduce the variance of cross-study estimates.

Transferability: The validity of extending an inference about an estimand to different sampling units or a different population of units. Transferability could be measured by the accuracy of a predicted estimand for a target population or observation, quantified by the difference between the transferred estimand and the 'true' estimand.

Weighting: Considered a hallmark of formal meta-analysis, the precision-weighting of each effect size by the inverse of its variance ensures that more precise studies make a larger contribution to the meta-estimate. Weighting serves only to increase the precision of the meta-estimate and the power of tests, not the accuracy of meta-estimation[96]. In the presence of bias, it can lead to precisely wrong estimates[30].

Figure legends

Figure 1. Generality - which we use synonymously with external validity - comprises the generalisability and transferability of estimands drawn from primary and synthetic research. Syntheses collate data from primary studies, each of which usually has a well-defined and narrow context relative to the context of the synthesis, and these studies are here each represented by a fruit of one of several types. Collated, these studies form a reference sample from a hypothetical population of studies, which together cover a broader context (here of fruits, either implicitly or explicitly defined by the researcher). Generalisability concerns the validity of an inference based on a sample that is randomly or non-randomly drawn from the target population (left column). Transferability concerns the validity of inferences based on a reference sample, when applied to either a different target population or unit (target context). Transfer across space is shown as an example, to sites in a different spatial location (middle row), or an individual target site from a different population (bottom row), which may also differ in temporal or taxonomic context to the reference sample. In both cases, the synthesised samples and the populations may have well-defined or poorly defined contexts. Here, the context of the synthesis is represented by the distribution of individual studies (fruits) within three measured or unmeasured dimensions of parameter space, e.g. edaphic, taxonomic, climatic variables (*V*) that vary depending on context and may influence the outcome of a study. In our example, the hypothetical reference and target contexts overlap (within the parameter space shaded blue) despite being on different continents.

Figure 2. Six steps to transfer an estimand to a target context

Figure legend in Box 1:

Two approaches used to synthesise primary studies, represented as different fruits, that have measured responses of some ecological variable *Y* to variable *X* and effect modification by variable *Z*. See Table 1.

Figure legend in Box 2:

Left: Effect sizes representing the effects of plantation thinning on abundance of saplings and seedlings depend on thinning intensity, showing grey-shaded 95% CI in the regression based on between-study and within-study uncertainty; values above horizontal dotted line signify higher abundance in thinned than unthinned stands. Point colour and shape combinations correspond to study identifiers, while point size is proportional to estimated weights. Middle: Spatial distribution of study sites in Japan. Right: Distribution of studies in parameter space according to mean annual rainfall and elevation. Coloured study locations are overlain on parameter space occupied by plots dominated by sugi or hinoki surveyed in a national forest inventory[89] (grey shading corresponds to plot density; see Appendix S2 for details).

Figure legend in Box 3:

Selection diagram approach for the effect of forest thinning on understory biodiversity (Adapted from Pearl & Bareinboim (2013)[75]