**Implications of scale dependence for cross-study syntheses of biodiversity differences**

Rebecca Spake1,2, Akira S. Mori3, Michael Beckmann4, Philip A. Martin5,6, Alec P. Christie6, Marlyse C. Duguid7, C. Patrick Doncaster2

**1** School of Geography and Environmental Science, University of Southampton, Highfield Campus, Southampton, SO17 1BJ, U.K. R.Spake@soton.ac.uk, <https://orcid.org/0000-0003-4671-22255>

**2** School of Biological Sciences, University of Southampton, Highfield Campus, Southampton, SO17 1BJ, U.K. cpd@soton.ac.uk, <https://orcid.org/0000-0001-9406-0693>

**3** Graduate School of Environment and Information Sciences, Yokohama National University, 79-7 Tokiwadai, Hodogaya, Yokohama, 240-8501, Japan. akira.s.mori@gmail.com, <https://orcid.org/0000-0002-8422-1198>

**4** UFZ – Helmholtz Centre for Environmental Research, Department of Computational Landscape Ecology, Permoserstraße 15, 04318 Leipzig, Germany. michael.beckmann@ufz.de, [https://orcid.org/0000-0002-5678-265X](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Forcid.org%2F0000-0002-5678-265X&data=01%7C01%7CR.Spake%40soton.ac.uk%7C19b4d8f8af3343d74a9008d81834ed07%7C4a5378f929f44d3ebe89669d03ada9d8%7C0&sdata=6IuVDPK0MgYbkAtBudaPZH%2Bf4L7Wv1DjIwtN7VAkKHA%3D&reserved=0)

**5** BioRISC (Biosecurity Research Initiative at St Catharine’s), St Catharine’s College, Cambridge CB2 1RL, UK. pam79@cam.ac.uk, [http://orcid.org/0000-0002-5346-8868](https://eur03.safelinks.protection.outlook.com/?url=http%3A%2F%2Forcid.org%2F0000-0002-5346-8868&data=01%7C01%7CR.Spake%40soton.ac.uk%7C043c2c0143e3429f64c108d818299f0e%7C4a5378f929f44d3ebe89669d03ada9d8%7C0&sdata=JZ4VrnoyhUd6xb1oufefOVHz3v6dkbNF1ym6SBn1h1o%3D&reserved=0)

**6** Conservation Science Group, Department of Zoology, University of Cambridge, The David Attenborough Building, Downing Street, Cambridge CB3 3QZ, UK. apc58@cam.ac.uk, <https://orcid.org/0000-0002-8465-8410>

**7** Yale School of the Environment, 195 Prospect St, New Haven, CT 06511, U.S.A. marlyse.duguid@yale.edu, <https://orcid.org/0000-0003-1465-0947>

***Running title:*** Scale-dependent meta-analysis

***Keywords***: accuracy, biodiversity, effect size, grain, meta-analysis, multilevel model, precision, scale, synthesis

***Article type****:* Reviews and Syntheses

***Article length***:Abstract: 195 words, Main text: 7,645 words (7,461 excluding headings); 1 table, 8 figures, 81 references.

***Corresponding author****:* Rebecca Spake, School of Geography and Environmental Science, University of Southampton, Southampton, SO17 1BJ, U.K. R.Spake@soton.ac.uk

***Author contributions***: RS conceived the idea, conducted the analyses and wrote the first draft. CPD developed the conceptual idea with RS and contributed to the first draft. MD supplied empirical data. All authors contributed substantially to interpretation and editing of the manuscript.

***Data accessibility statement***: The Supporting R Scripts for generating the simulated data, and the empirical data from Duguid & Ashton are available at: <https://doi.org/10.6084/m9.figshare.12496955>. Empirical data from the meta-analysis of Powell, Chase and Knight (2011) are available in an online file at <http://www.amjbot.org/cgi/content/full/ajb.1000402/DC1>.

### Abstract

Biodiversity studies are sensitive to well-recognised temporal and spatial scale dependencies. Cross-study syntheses may inflate these influences by collating studies that vary widely in the numbers and sizes of sampling plots. Here we evaluate sources of inaccuracy and imprecision in study-level and cross-study estimates of biodiversity differences, caused by within-study grain and sample sizes, biodiversity measure, and choice of effect-size metric. Samples from simulated communities of old-growth and secondary forests demonstrated influences of all these parameters on the accuracy and precision of cross-study effect sizes. In cross-study synthesis by formal meta-analysis, the metric of log response ratio applied to measures of species richness yielded better accuracy than the commonly used Hedges’ *g* metric on species density, which dangerously combined higher precision with persistent bias. Full-data analyses of the raw plot-scale data using multilevel models were also susceptible to scale-dependent bias. We demonstrate the challenge of detecting scale dependence in cross-study synthesis, due to ubiquitous covariation between replication, variance and plot size. We propose solutions for diagnosing and minimising bias. We urge that empirical studies publish raw data to allow evaluation of covariation in cross-study syntheses, and we recommend against using Hedges’ *g* in biodiversity meta-analyses.

## Introduction

The global scale of biodiversity loss amounts to a sixth mass extinction event in Earth’s history that demands transformative change in human behaviour (Ceballos *et al.* 2017; IPBES 2019). Efforts to curb extinctions are principally implemented at the local scale, by restoring, conserving and sustainably managing ecosystems, and can be effective only when informed by scientifically credible evidence (Sutherland *et al.* 2004). However, individual field studies of local-scale drivers inevitably report idiosyncratic effects on biodiversity, hindering efforts to generalise and effectively transfer conservation actions beyond case studies (Spake *et al.* 2019a). Idiosyncrasies arise amongst individual studies from the impossibility of simultaneously maximising realism, precision and generality in study design (Levins 1966) when logistical constraints limit taxonomic and biogeographic scope.

Generality can be increased by cross-study synthesis of individual study outcomes, and analysis of their variation across, for example, biomes, time periods, ecosystems or taxa (Gurevitch *et al.* 2018). However, contributing field studies typically vary widely in scale (Keil & Chase 2019), which can offset the gains in generality with costs in realism (Figure 1). The problem of scale has long been recognised as a central issue in ecology (Levin 1992), yet the effects of scale in biasing biodiversity responses defy straightforward analysis (Chase *et al.* 2019). Although numerous case studies have demonstrated scale-dependence in biodiversity responses to local drivers (e.g. Hill & Hamer 2004; Chase *et al.* 2018), cross-study syntheses have rarely addressed scale influences (but see Daskalova *et al.* (2020) for a quantification of landscape-scale influences on local biodiversity change). Here we aim to measure relative magnitudes of scale-dependence, in both meta-analysis of study-level summary statistics and cross-study raw data analysis, with the goal of providing guidelines for minimising scale-dependent bias (hereon ‘scale bias’).

We first review sources of scale-dependence within individual field studies, present in summary statistics of biodiversity differences between control and treatment groups. We then test for an influence of these sources, by simulating sets of biodiversity studies comparing old-growth to secondary forests, and subject them to both meta-analysis and full-data analysis. We go on to seek empirical evidence of the issues revealed by the simulations, by revisiting two published meta-analyses, measuring biodiversity responses to forest management (Duguid & Ashton, 2013) and to plant invasion (Powell et al. 2011). We provide recommendations on how to address scale dependence in future syntheses.

## Scale dependence in Summary Statistics from field studies

Measures of biodiversity from any single community increase with increasing sampling effort and spatial grain of sampling, in a non-linear relationship that depends on the total numbers of species, the total and relative abundances of individuals, and their spatial distribution (Chase & Knight 2013). In formal meta-analysis, the outcome of each component study is expressed as an effect size, typically representing the mean difference between two treatment groups (Koricheva, Gurevitch, & Mengersen, 2013). Effect sizes and their variances are estimated from summary statistics: means, variances and sample sizes from control and treatment groups. For a response variable of biodiversity, and groups that represent different land-use types or points in time, these statistics are all susceptible to vary with the scale of the study, as defined by its sampling grain, in ways that we detail below (summarised in Figure 2A).

### Scale dependence in mean biodiversity differences

Most biodiversity comparisons use measures of species density, given by the number of species per unit area or per sample, adjusted for sampling effort (Gotelli & Colwell, 2001). Interrogation of species’ abundances in standardised samples can reveal whether observed differences in density are due to differences in species abundance distributions or total numbers of individuals (e.g., McGlinn *et al.* 2018; Shimadzu, 2018). Nevertheless, cross-study syntheses typically compare differences in species density without reference to abundance distributions. Even if syntheses equalise sampling effort between treatments within studies (e.g. using sample-based rarefaction, e.g. Blowes *et al.* 2019; Antão *et al.* 2020; Daskalova *et al.* 2020), an implicit assumption remains: that differences observed at one grain would be observed at others. This is despite the fact that empirical studies have demonstrated that species density responses to land use, invasive species, and grazers/predators can vary in magnitude and even direction with grain size (Chase *et al.*, 2018). In contrast, biodiversity measures that use abundance distributions to extrapolate communities towards asymptotic richness (e.g. Chao-1, Chao 1984), are considered robust to variations in grain size if sample coverage is sufﬁciently large (Hortal *et al*. 2006).

### Scale dependence in study variance

Any change in the scale of measurement of a variable also changes the variance of that variable (Wiens 1989). Sampling grain can therefore influence the variability among plots, for a given level of replication. For example, smaller plot sizes can result in fewer individuals being sampled per plot, and consequently higher within-treatment variation in observed taxonomic diversity (Connell & Sousa 1983; Chase & Knight 2013), in addition to greater beta diversity among plots (Sreekar *et al.* 2018). Scale dependence of variance is likely to be strongest in non-uniform environments with patchy distributions of individuals and species (Král *et al.* 2010; Avery & Burkhart 2015; Grussu *et al.* 2016). This high variability among small plots has the potential to weaken observed relationships between environmental gradients and biodiversity (Field *et al.* 2009).

### Scale dependence in replication

Sample size often varies among studies in a trade-off with the spatial grain of sampling, caused by the logistical, cost and time constraints inherent to field research (Monserud, 2004; Ramage et al., 2013). As a consequence, ecological meta-analyses routinely collate studies with many-fold differences in replication (Doncaster & Spake 2018). Whilst replication boosts the signal-to-noise ratio in ecological data and increases precision, it does not by itself influence the proportion of the signal caused by bias (De Palma *et al.* 2018; Christie *et al.* 2019). For example, a study with small plot sizes may compensate for a higher inherent variability by expanding sample size to increase statistical power (Rosenberg *et al.* 2013). When included in a meta-analysis, the small plots sizes expose the meta-estimate to bias (Chase & Knight 2013; Chase *et al.* 2018), which is then amplified by the large sample size conferring a high precision-weighting on the study. This issue may become especially problematic in meta-analysis if replication correlates negatively with plot size across studies.

## Scale dependence in Cross-Study syntheses of biodiversity differences

Cross-study synthesis usually involves one of two approaches: meta-analysis or full-data analysis. Meta-analysis typically combines effect sizes using a standard meta-analytic (intercept-only) model, to estimate an overall effect-size mean (meta-estimate), and heterogeneity in study outcomes. Treatment-covariate interactions can be investigated using meta-regression, which quantifies associations between study-level treatment effects and study-level covariates, such as biome or taxonomic group (Nakagawa *et al.* 2017 & Gurevitch *et al.* 2018). Full-data analyses, in contrast, typically synthesise the raw biodiversity data across all studies using multilevel models (also known as hierarchical or mixed-effects models). Study-level random effects can be incorporated to allow the biodiversity response to vary among studies. Overall treatment effects correspond to fixed, population-level estimates.

### Standardising study-level effect sizes against within-study variance can amplify scale bias

In ecological meta-analysis, the two most commonly used metrics of meta-effect size are the standardised mean difference, estimated as Hedges’ *g* (Hedges, 1981; hereafter *g*) and the log response ratio (Hedges et al., 1999; hereafter *LR*). Hedges’ *g* standardises the mean difference between treatment and control groups by their pooled standard deviation, to create a metric of difference per unit of residual variation; *LR* standardises the mean of the treatment group against the mean of the control group to create a ratio (formulae in Table 1). Hedges’ *g* therefore confounds within-study variation with the observed mean difference (e.g. Monserud, 2004; Osenberg, Sarnelle, & Cooper, 1997; Spake & Doncaster, 2017). This confounding is important in the context of ecological meta-analyses, because if the pooled standard deviation depends on factors such as sampling grain, and if these factors co-vary with taxonomic group and habitat, patterns in *g* may emerge that simply reflect differences across studies in within-study variation (Osenberg *et al.* 1997). Moreover, the standardised mean difference itself contributes to the estimate of its variance (Table 1), meaning that any bias in estimation of the true *g* will propagate into bias in the estimated variance.

Although Hedges (1982) proposed an alternative, ‘unbiased’ variance estimator that does not incorporatethe standardised mean difference(Table 1), few meta-analyses have used it (Hamman *et al.*, 2018). Gurevitch, Curtis, and Jones (2001) pointed out that a ratio-based metric such as *LR* avoids both these issues, of the standard deviation contributing to the effect size, and the effect size contributing to the variance. Nevertheless, *g* and its conventionally estimated variance remains a commonly-used metric in meta-analyses of biodiversity change, typically justified by the presence of zero mean values that preclude the estimation of *LR*.

### Precision-weighting effect sizes can amplify scale bias

The hallmark of formal meta-analysis is precision-weighting of effect sizes (Gurevitch, et al., 2018), such that more precise studies make a larger contribution to the meta-estimate (Hedges & Olkin 1985; Gurevitch & Hedges 1999). Weighting serves only to increase the precision of the meta-estimate and the power of tests, not the accuracy of meta-estimation (Gurevitch & Hedges 1999). A precise estimate size has narrow confidence limits, while an accurate one lies close to the true value for the sampled population. A metric is unbiased when accuracy increases with precision. A metric that is precise but biased is much more undesirable than one that is imprecise but unbiased, because the high precision appears to suggest high accuracy whilst actually yielding high inaccuracy (Figure 2B). Any bias in study-level variance will in turn bias the conventional precision-weighting of each study given by the inverse of its variance (Hedges, 1983). The resulting gain in precision from weighting may therefore come at a cost in accuracy.

### Full-data analyses are also subject to scale biases

Researchers increasingly have access to raw, site-level biodiversity data, following calls for an ‘open synthesis community’ (Naka*gawa et al.* 2020). Full-data analyses are considered important for resolving issues regarding study-specific analytical designs, and are a ‘gold standard’ in medicine (Mengersen *et al.* 2013). Previous full-data analyses of biodiversity differences have yielded population means of fixed treatment effects from multilevel models that incorporate group-level (i.e., study-level) random effects (e.g. Newbold *et al.* 2015; Jung *et al.* 2019). A fundamental property of multilevel models is ‘borrowing strength’, wherein individual group 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 (Clark, 2019). Like weighting in formal meta-analyses of effect sizes, shrinkage estimation functions to reduce the variance of cross-study estimates. It follows that the degree of study-level shrinkage is also determined by study-level variance, and may have similar sensitivity to scale-bias amplification.

### Incorporating scale-dependence in cross-study syntheses will be challenging

Given that biodiversity differences observed within studies depend on the spatial grain of sampling, any cross-study syntheses of biodiversity differences will be greatly enhanced by an explicit consideration of scale (Chase et al. 2019). Quantitative syntheses may factor in this known source of variability by including the sampling grain size as a covariate or ‘effect modifier’ in a cross-study regression, or they may seek to identify spatial grains at which a treatment has a particular conditional effect. In either case, the detection of scale dependence serves to warn against inappropriate generalisation across scales. Nevertheless, it remains unclear how straightforward it will be to incorporate scale into future cross-study syntheses (Chase et al. 2019). Detection may prove challenging, for example, because the influence of spatial grain on effect size may be obscured by scale-dependent sample means, variances and replication (Figure 2C).

## Simulation of biodiversity differences

In order to quantify the influence of sampling grain and replication on precision and accuracy of study-level and cross-study effects, we simulated empirical studies that compare tree biodiversity between old-growth and secondary forests. These two forest types represent two treatment levels (control and treated, respectively) that differ in the number of species. For each study, we randomly sampled *N* forest patches of each type from a set of forest patches that defines the study scope. We then randomly sampled tree communities within the patches using square plots of fractional size *A*, and estimated biodiversity with two measures: species density and asymptotic richness (Table 1). We calculated effect sizes from metrics that quantify the influence of forest type on biodiversity, measured as plot density or species richness (Figure 3). We then obtained cross-study estimates of the sizes of meta-effects from meta-analysis, and population effects from full-data analysis, for a hypothesised dependence of biodiversity on forest type. The simulation thereby mimicked analyses that estimate cross-study effects without considering variation in sampling grain across studies. We demonstrate varying degrees of scale dependence in the cross-study syntheses, depending on choice of biodiversity measure, effect-size metric, weighting scheme, variance estimator, and the degree of trade-off between replication and grain.

### Generation of sampled patches and true species richness

We set the true species richness of forest patches (*PatchRich*T) to be higher on average for old-growth than for secondary forest (e.g. Gibson et al., 2011; Newbold et al., 2015; Spake, Ezard, Martin, Newton, & Doncaster, 2015; Spake, Yanou, Yamaura, & Kawamura, 2019b). For each study, we created populations of 100 old-growth and 100 secondary forest patches, each having a true patch species richness *PatchRich*T drawn from a normal distribution with standard deviation σ = 10, and means of  = 120 species for old-growth and 70 species for secondary (Figure 3.1). All forest patches had the same patch area (rationale in Appendix S1, Figure S1.1).

### Plot sampling of forest patch communities

For each forest type, we selected *N* patches from the population of 100 patches of that type, where *N* took values of 6, 12, 18, 24 and 30 replicate patches to assess the effect of replication. In order to avoid the complication of nested replication, we sampled the community with one plot per patch. The forest patch thus corresponds to the unit of effective replication in our setup, as it would in statistical analyses of actual studies with or without replicate plots per patch.

We used R version 3.5.3 (R Core Team 2019) and the R package mobsim (May *et al.* 2018) to simulate the tree communities of each sampled patch. This package allows manipulation of the total number of individuals, the species-abundance distribution (SAD) and the degree of intraspecific spatial aggregation (May *et al.* 2018). Chase and Knight (2013) showed that any factor (such as forest type) that influences the total richness or abundance, SAD and/or spatial aggregation of species within a community, will shift the overall shape of its species accumulation curve (SAC), such that SACs from different communities will not be parallel to each other. Consequently, effect sizes based on biodiversity differences are scale dependent, even when sampling is standardised by area or the number of individuals. For simplicity and conservative testing, we varied only the species richness between old-growth and secondary forest communities, whilst holding both forest types to the same total abundance, SAD and degree of spatial aggregation. We evaluate the implications of this setup in the discussion. We used the mobsim function sim\_poisson\_community() to simulate communities with individuals randomly distributed across space, i.e. with no spatial aggregation. We simulated communities of 2000 tree individuals (stems) per patch following a log-normal SAD, with a species richness corresponding to the particular *PatchRich*T sampled from the population of patches.

We sampled each simulated forest community using a square plot placed randomly within the square forest patch using the mobsim function sample\_quadrats(). We defined sampling grain, *A*, as the proportion of the forest patch area sampled by the plot, setting 10 values of *A*, at 0.05, 0.1, 0.2, …, 0.9. One hundred communities were created for each combination of *N* patches × *A* plot area, yielding 5×10×100 = 5,000 simulations.

### Biodiversity measures

We used plot-level data to derive two alternative measures of alpha diversity. The observed number of species per plot, *PlotDensity*E, estimates the true *PatchRich*T per plot area; the estimated number of species per patch, *PatchRich*E, estimates the true *PatchRich*T (Figure 3). These are the two most commonly used measures of biodiversity in empirical studies and meta-analyses (Magurran, 2004); confusingly, the literature does not always distinguish clearly between them in references to ‘species richness’. For *PatchRich*E, we applied the Chao-1 non-parametric richness estimator (Chao 1984) in R package vegan v.2.5-6 (Oksanen *et al.* 2019). This method uses information on the rare species in an assemblage to adjust for the number of species present but undetected, as a robust estimate of the minimum number of species in the patch (Gotelli & Colwell, 2009).

The selected properties of the community described above generated area-based species accumulation curves typical for these forest types in both empirical and simulation studies (Figure S1.2; e.g. Guariguata *et al.* 1997; Jiménez *et al.* 2016). Area-based standardisation (using *PlotDensity*E) typically results in greater under-sampling of richer habitat (here old-growth forest), leading to underestimated effect sizes (Gardner *et al.* 2007; Chase & Knight 2013; Curran *et al.* 2014).

### Effect-size metrics and their variances

For each simulated study, which corresponded to a particular replication × grain (*N* × *A*) combination, we calculated the mean and standard deviation of each biodiversity measure (*PatchRich*Eand *PlotDensity*E) from the *N* replicate patches of each forest type. These measures were used to calculate the alternative effect-size metrics *g* and *LR*, and their variances. For *g,* we used the conventional variance (*Vg*), and also Hedges’ (1982) alternative (*Vg\_*alt; Table 1), with the principal difference being that the formula for *Vg*  containsthe standardised mean difference*,* whereas the formula for *Vg\_alt* is independent of it. Prior to calculation of *LR*, we ensured that all of the control and treatment means satisfied Geary’s test, a diagnostic for vulnerability to biases caused by low replication, large differences between treatment and control means, and/or near-zero means (Geary, 1930; Lajeunesse, 2015).

### Accuracy and precision of study-level effect sizes

Accuracy is the distance of a sample estimate from the population value it estimates, with a small distance signifying high accuracy. For the simulation, we used a standardised accuracy of effect size = [estimated – population] / population. Accuracies of study-level effect sizes were calculated using the true, population-level, average richness values of each study (*PatchRich*T = 120 for old-growth, and 70 for secondary; Table 1). The resulting proportion allowed direct comparisons of accuracy between *g* and *LR*. This accuracy metric encompasses two types of sampling error: (i) random error incurred by sampling *N* patches from each of the populations of 100 patches, and (ii) systematic error arising from the estimation of patch-level biodiversity from plot-level data. Here we leave them aggregated for the purpose of analytical concision, given that sample-level effects aim to estimate population-level effects. Moreover, their aggregation allows us to examine the relationship of accuracy to the degree of collinearity between *A* and *N*, which influences both types of error.

Precision is the distribution of replicate estimates around their mean, with a tight distribution signifying high precision. We quantified precision as the inverse of effect-size variance, following the convention of this metric being used for precision-weighting of studies in meta-analyses. We examined the relationship of accuracy to precision (Figure 2B), and how it may change with sampling grain, replication, effect-size metric and biodiversity measure.

### Meta-analyses of simulated studies and meta-regression on plot size

To evaluate the influence of a trade-off between grain and replication, we tested for a relationship of meta-estimate magnitude to the strength of collinearity between *A* and *N* across study pools. To this end, we generated 150 study pools, spread over a spectrum from strong to negligible negative correlation of *A* with *N*. We used function genCorOrdCat() in R package simstudy (Goldfeld 2019) to generate correlation coefficients between -0.80 and +0.15 from the simulated ranges of *A* (0.05 to 0.9) and *N* (6 to 30). For each study, we estimated biodiversity in the communities of *N* forest patches sampled from each of the populations of 100 old-growth and 100 secondary patches, following the procedures described above.

We meta-analysed the study-level effect sizes of all 150 studies in each pool with R package metafor v.2.4-0 (Viechtbauer 2010), to obtain the overall meta-estimate of effect size for dependence of biodiversity on forest type. This was done for each effect-size metric (*g* and *LR*) derived from each biodiversity measure (*PlotDensity*E and *PatchRich*E).

We explored the effect of alternative weighting schemes on meta-estimate accuracy. The method of incorporating inverse-variance weights differs between fixed- and random-effect meta-analyses (Table 1). Precision-weighting used the weight 1/(*Vi* + τ*2*), where *Vi* is the variance in *g* or in *LR* for the *i*-th study, and τ2 is the among-study variance estimated by restricted maximum likelihood. This yielded meta-estimates of random effects (assigning τ2 > 0) or fixed effects (assigning τ2 = 0). We conducted meta-analyses of *g* using both variance estimates *Vg* and *Vg\_*alt.

We first checked for a known bias in meta-analysis of little-replicated studies, due to random error in estimation of the true study-level variances accumulating in the estimates of meta-effect and meta-variance (Doncaster & Spake, 2018). This bias is most pronounced for meta-analysis with Hedges’ *g* of studies having predominantly *N* ≤ 6 replicates. For our simulations with *N* ranging between 6 and 30, the mean-adjustment to *Vg* that corrects this bias made negligible difference to its magnitude (Figure S1.3). We therefore used conventional calculations of *V*, for compatibility with most previous meta-analyses.

We quantified the relationship of meta-estimation accuracy to weighting scheme, effect-size metric, biodiversity measure and degree of collinearity between *A* and *N*. We further conducted meta-regressions of effect size against log(*A*) for study pools with uncorrelated grainand replication*,* to test for any influence on the detection of trend and gradient of regression slope due to the weighting scheme and effect-size metric. For each meta-regression, we estimated the regression coefficient (gradient of slope) and its significance with function rma.mv() in R package metafor. Effect sizes were converted to *z-*scores (centred, with standard deviation of unity) prior to modelling, to enable direct comparisons of meta-regression coefficients between models of *g* and *LR*.

### Full-data analysis of simulated studies

We analysed the raw, plot-level data from each study using R package lme4 v.1.1-23 (Bates *et al*. 2012). A generalised linear mixed model was fitted to *PlotDensity*E (count data) with a Poisson error and log link function; a linear mixed model was fitted to *PatchRich*E (non-integer values) with a normal error and identify link function. For estimation of the population-mean effect of forest type across study pools (akin to an intercept-only formal meta-analysis), we included forest type as a fixed explanatory variable (old-growth [0] or secondary [1]), specifying random intercepts and random slopes of forest type for each study. We further fitted models that included interactive effects of forest type and sampling grain [log(*A*)] (akin to meta-regressions of effect size on grain), with random intercepts and slopes for forest type. To visualise the interactive effect of forest type and log(*A*) as estimated from the full-data analyses, and compare this to the meta-regressions, we estimated the effect of forest type, conditional on values of *A* using R package interplot v.0.2.2 (Solt *et al*. 2018). Further details of multilevel model specifications are available in Appendix S2. To aid the visualisation of results, we present outputs from meta-analyses in blue or yellow and outputs from full-data analyses in red.

## Simulation results

### Relationship of accuracy to precision at the study level

For the calculation of *g* with conventional variance *Vg*, a negative relation of accuracy to precision became increasingly shallow with increasing replication (*N*,as shown by the separating lines in Figure 4a). The most overestimated effect sizes were mainly associated with larger grain size *A*, and they universally had the lowest precision regardless of *N*. The most underestimated effect sizes were associated with smaller grain sizes, and they universally had the highest precision for a given *N*, making them precisely wrong, and more precisely wrong for higher *N*. This strongly defined bias reflects the fact that *g* itself is used in the calculation of *Vg* and hence precision. In precision-weighted meta-analysis, it will elevate the contribution of underestimates of effect size and suppress the contribution of overestimates.

For the alternative calculation of *g* with *Vg\_*alt, the removal ofstandardised mean differencefrom the variance calculation removed the negative dependency of accuracy on precision (Fig. 4b). Precision now depends only on *N*, the sole parameter defining *Vg\_*alt.

Effect sizes for *LR* derived from *PlotDensity*E tended to be underestimated for smaller grain sizes regardless of replication; accuracy had no systematic trend with precision (Fig. 4c). Effect sizes for *LR* derived from *PatchRich*E had the most desirable properties for meta-analysis, given by a funnel-shaped relationship such that accuracy increases with precision (Fig. 4c, -right, cf. Figure 2b). Strongly over- or under-estimated effect sizes had low precision, suggesting therefore that they would contribute little to precision-weighted meta-analyses. High precision associated with high accuracy, and therefore little bias. Over- and under-estimation were equally prevalent, and precision increased with both grain and replication.

For study means of plot-level biodiversity (relevant to full-data analyses), underestimation with *PlotDensity*E generally declined at larger grain sizes (Figure 4d). *PlotDensity*E was underestimated more for old-growth than secondary forests (Figure 4d & Figure S1.2). While for a given grain size, a funnel shape is evident, a slightly negative trend of accuracy to precision arises across all plot sizes, because small plots with low mean values of *PlotDensity*E correspond to low variance and so high precision. *PatchRich*Ebalanced over- and under-estimation even at relatively small grains, albeit with lower precision.

### Cross-study effects with grain-by-replication trade-offs

Meta-effects were universally underestimated by meta-analysis, due to their inclusion of studies with small *A*. Negligible bias was found only for *LR* derived from *PatchRich*E (Figure 5a-c, yellow dots in right-hand graphs). The minimal bias shown here results from the efficiency of the Chao-1 estimator, even at small grains. Weighting had little effect on the accuracy of meta-estimates of *LR* derived from either measure of biodiversity. As predicted, weighting decreased the accuracy of meta-estimates of *g*, and more with *Vg* than with *Vg*\_alt. Stronger negative correlations of *N* with *A* strengthened the underestimation of meta-effects from weighted meta-analysis, although this trend was only sharply defined in fixed-effects meta-analysis of *g* (Figure 5b, blue dots).

Global meta-estimates of *g* were more accurate than *LR* for unweighted estimates derived from *PlotDensity*E. Although this might seem to reflect value in *g,* the meta-level accuracy arises only from an approximate balancing out in our simulations of predictable overestimation at large *A* and underestimation at small grains (Figure 5a-c, and corroborated by additional analyses presented in Appendix S3). Empirical data are unlikely to achieve this exact balancing out, however, as study pools typically do not have uniform distributions of plot size and replication (e.g. Duguid & Ashton, 2013), or are not capable of sampling high fractions of patch area. In further simulations, *g* lost accuracy with skews towards greaterreplicationor larger grain size (Appendix S3). In all simulations with either metric, *PlotDensity*E yielded lower accuracy than *PatchRich*E with *LR* (Figure 5a-c comparing columns, Appendix S3). For both biodiversity measures, the effect of weighting on meta-estimate accuracies of *g* was stronger with *Vg*rather than *Vg*\_alt, due to the greater weight given to studies with higher precision and therefore more underestimation of *g.* Thisresulted in meta-estimates having lower accuracy with *g* weighted by *Vg*\_alt, than with *LR* (Figure 5a).

For population means from full-data analyses, accuracy showed a similar trend to meta-analysis using *LR*, with greater accuracy for models of *PatchRich*E than *PlotDensity*E (Figure 5d). *PatchRich*E nevertheless yielded less accurate population means in full-data analysis than meta-estimates using *LR* in meta-analysis. The lower accuracy results from shrinkage, wherein group (study) estimates are ‘shrunk’ toward the population mean as a function of the relative variance of each estimate (Appendix S4).

Cross-study regressions of treatment-by-grain effects showed *LR* and full-data analysis to be better discriminators of scale effects than *g* (Figure 6a-d, Figure S5.1). The relationship of *g* with *A* was generally more strongly negative for *PatchRich*E than *PlotDensity*E, and stronger when precision-weighted by *Vg* than *Vg*\_alt (see also Figures S5.1 and S5.2). In contrast, both full-data analysis and meta-regression with *LR* appropriately revealed stronger scale dependence of forest type effects on *PlotDensity*E than *PatchRich*E (as shown by stronger slopes for *PlotDensity*E in Figure 6c-d). For example, at grain sizes covering 5% of the forest patch, the full-data analysis of *PlotDensity*E estimated a conditional forest type effect of approximately -0.30, which strengthened to -0.54 at grain sizes covering 75% (Figure 6d left). The type and method of cross-study synthesis least influenced by grain size is meta-analysis with *LR* using *PatchRich*E, as shown by the predicted values in Figure 6c deviating the least from the true effect size.

## Empirical meta-analyses of biodiversity change

We explored variation in the detectability of scale dependence according to effect size metric and weighting scheme, by meta-analysing studies retrieved from two published syntheses of biodiversity change. We ran meta-regressions of effect size as a response to grain size.

Duguid and Ashton (2013) synthesised summary statistics from studies measuring differences in plant biodiversity between unmanaged (control) and logged (treatment) forest. We used the grain sizes (quadrat areas) from these studies to test for scale dependence and potential implications for meta-estimation. We restricted our re-analysis to studies of even-aged treatment forests at mid to late successional stages as classified by Duguid & Ashton (2013), which included 63 comparisons from 24 publications. All studies compared speciesdensity measured in fixed-area plots ranging in size between 1 and 1000 m2. Management treatments included timber harvesting by clearcutting and shelterwood cutting, which may reduce microsite heterogeneity (Duguid & Ashton, 2013). We therefore expected management to reduce species density, resulting in strongly negative effect sizes, at least for studies measuring density at relatively small sampling grains that encompass homogeneous microsites. We expected effects to become less negative with increasing grain size, as larger quadrats begin to encompass multiple microsite types with distinct communities.

Powell, Chase, and Knight (2011) meta-analysed effects of invasive plants on species richness of invaded communities across quadrats ranging in size between 0.09 and 2500 m2. Their synthesis used summary statistics on 123 studies from 56 publications measuring species density differences between uninvaded (control) and invaded (treatment) plots. They detected a relationship between grain size and effect size, which was stronger for *LR* than for *g*. The effect of invasion was more strongly negative at smaller grains, and weaker at larger grains, due to invaders reducing the occupancy of common species to a greater degree than rare species.

We conducted formal meta-analyses to estimate overall meta-effects of forest management (Duguid and Ashton, 2013) and invasion (Powell, Chase & Knight, 2011) on plant species density, for random- and fixed-effects weighting and no weighting. We specified control groups nested in publication identifier as a random factor, to account for multiple treatments compared with the same control group within publications. We further ran meta-regressions with log(grain size) as a covariate, using the same random structure, for each weighting scheme, and the conventional variance estimators for *g* and *LR* (*Vg* and *VLR*), from which we estimated coefficients for grain size (i.e. regression slopes). Effect sizes were converted to *z-*scores prior to modelling, to enable direct comparisons of meta-regression coefficients between models of *g* and *LR*.

## Empirical results

### Meta-analyses of forest management effects on plant species density

Within-study replication was higher for studies with smaller grain sizes (Spearman’s  = -0.71, Figure S6.1). Both *g* and *LR* showed a tendency to increase with *A*, ( = 0.50 and 0.46, respectively), mostly switching from negative to positive across the plot-size gradient (Figure S6.1). Variance in *g (Vg)* and *LR* also increased with *A* ( = 0.60 and 0.30 respectively, Figure S6.1).

Meta-analyses detected no overall effect of forest management on plant species density across the study pool (all meta-estimate CIs encompassed zero, Figure S6.2). Meta-regressions nevertheless revealed an influence of grain size on effect-size estimates for some combinations of weighting scheme and effect-size metrics (Figure 7; Figure S6.3). Grain size had a detectable and positive influence on effect sizes for random-effects models of both *g* and *LR,* and for unweighted *LR* (Fig. 7, top row and bottom right; Figure S6.3). Random-effects and unweighted meta-regression of *LR* yield similar results because high among-study variance (τ2) will pull random-effects meta-analysis of *LR* towards the character of unweighted meta-analysis. The influence was strong enough to suggest a switch in the treatment effect from negative at small grain to positive at large grain. No such influence was detected in fixed-effect models, and unweighted *g*.

### Meta-analyses of invasion effects on plant species density

Within-study replication was again higher for studies with smaller grain sizes (Figure S7.1). As a consequence, meta-analyses of these studies similarly revealed variation in both meta-effects and meta-regressions of effect sizes on plot size (Figures S7.2-S7.5), according to effect-size metric and weighting scheme. Meta-analyses detected an overall negative effect of invasion on plant species density across all analyses (Figure S7.2). For meta-regressions, plot size exerting a detectable and positive influence on effect sizes for random-effects models of *g* and *LR*, and for unweighted *LR* (Figure S7.5; full results in Appendix S7).

## Discussion

Here we have demonstrated that cross-study syntheses amplify a type of within-study scale bias, which results from under-sampling of species richness in intrinsically richer habitats at small grains. We have shown how the scale-dependent impacts on cross-study estimation depend on key decisions in quantitative synthesis, concerning the choice of effect-size metric and the weighting scheme in meta-analyses, and model specification in full-data analyses. We have further shown how these choices determine the detection and estimated magnitude of scale dependence. We provide guidance below for treating scale-dependence in cross-study syntheses of biodiversity differences, and for appraising existing syntheses (summarised in Figure 8; further elaborated in Appendix S8).

### Sensitivity of cross-study estimation to scale bias depends on type and method of synthesis

The validity of interpreting meta-estimates from summary statistics has been contested on two grounds. Firstly, constituent effects likely comprise a biased sample of true population effects, with over- or underrepresentation of geographic regions, land-use types, and taxonomic groups (Gonzalez *et al.* 2016). Secondly, studies are too heterogeneous in method or results to be meaningfully combined without conditioning on covariates (Poole & Greenland 1999; Ioannidis *et al*. 2008). Nevertheless, summary effects are often reported without reference to heterogeneity in sampling grain, and interpreted as patch-level biodiversity differences. We have demonstrated how such meta-estimates vary in their sensitivity to bias.

In our simulations, Hedges’ *g* provided study-level and meta-level effect sizes that were consistently less accurate than those of the log response ratio for estimation of patch richness. Because of the random sampling error generated by sampling forest patches from a population, study-level effect sizes may both over- and under-estimate the population effect size. Even a symmetrically distributed sampling error will bias meta-estimates of any effect derived from *g* using inverse-variance weighting, because the weights are too high for studies that underestimate its magnitude (Hedges 1983; Hamman *et al.* 2018). When combined with a systematically low weighting for overestimates, the net result is an under-estimated meta-effect size from precision-weighted meta-analysis using *g*.

For Hedges’ *g* with its conventionally-estimated variance*,* increasing grain size drives declining precision, and a sharply defined switch in bias from under- to over-estimation. For *LR* in contrast, increasing grain does little to change the relationship between accuracy and precision except with *PatchRich*E, where it tends to increase precision and accuracy. Both metrics and both measures consistently undervalued effect sizes estimated from small plots, due to under-sampling of richer old-growth forest patches. This led to meta-estimates being increasingly underestimated for study pools with increasingly negative correlations between grain and replication; these study pools contain a higher proportion of individual effect sizes that are precisely wrong, rather than imprecisely right.

Full-data analyses of *PatchRich*E, with forest type as the sole explanatory variable, had inferior meta-effect accuracy to meta-analyses of *LR*, and a greater sensitivity to any skewing towards smaller plot sizes across studies (Figure S3.5). This was due to inappropriate shrinkage incurred by omission of the forest-type-by-grain size interaction; when making a random-effects assumption about a factor, their effects are not estimated independently and individual estimates are pulled toward the common mean. Any skew or bias in the study pool can therefore become inflated. Indeed, shrinkage is not ‘correct’ unless all true covariate effects, including interactions, are included in the fixed part of the model (Bell *et al.* 2019).

Our results cast a more general doubt on the validity of estimating and interpreting cross-study effects from study pools of widely varying grain. Our simulations showed that even for an otherwise homogeneous study pool comprising a single taxonomic group (here trees), cross-study estimates have questionable utility on their own. For both effect sizes in meta-analyses and random slopes in full-data analyses, biodiversity differences based on species density vary in magnitude with sampling grain. The interpretation of effect sizes – whether from individual studies or from cross-study synthesis – can therefore yield inferences of practicable value only when set in the context of a specific sampling grain.

### Detecting scale dependence using meta-regression and full-data analysis

Chase et al. (2019) argued that simply including the scale of observation as a covariate in cross-study syntheses of biodiversity differences would only yield interpretable results if a single process predominated (and thus a single direction of scale-dependence). Our simple simulation, which varied a single community-level component, the total species richness, and kept constant the patch-level total abundances and SADs, has revealed that scale dependence still defies straightforward characterisation.

For both simulated and empirical data, the detection and magnitude of grain-size influences differed according to effect-size metric and weighting scheme. Hedges’ *g* was less discriminating than *LR*, and fixed-effects less discriminating than unweighted and random-effect models. These meta-estimates depend most strongly on study variance, which co-varies with grain and replication, obscuring the scale dependence in effect size. Fletcher & Dixon (2012) found that unweighted meta-regressions outperformed weighted meta-regressions, because the former did not make use of potentially misleading information on precision, except when precision covaries with leverage in the regression. Although we cannot know the true effect size and sampling variance for the empirical studies, this covariation may contribute to the observed differences in detectability of a grain-size influence in our empirical meta-regressions.

The observation that scale dependence may go undetected, even when grain size is included as a covariate/effect modifier in a meta-regression, is cause for concern. Previous studies may have wrongly concluded scale-independent effects of an environmental driver, and conservation agencies may have interpreted results to mean that a particular intervention would have the same effect at any scale. Less-than-rigorous applications of meta-analytical methods have used standard linear regression methods to precision-weight effect sizes (as noted by Gurevitch et al., 2018). This approach has the character of a fixed-effects meta-analysis, because it typically ignores among-study variance, which we have found to perform poorly in detecting scale dependence within a heterogeneous study pool. These studies may have wrongly dismissed the possibility of scale-dependent magnitudes or even directions of effect.

The alternative possibility is that scale dependence may be detected only as an artefact of effect-size metric and weighting scheme. Our simulations all had the same scale of true effect. As expected, meta-regressions of *LR* on grain size yielded only a weak influence of grain size on effect sizes derived from asymptotic species richness, because the Chao-1 estimator accurately estimates species richness, even at small sampling grains (Figure S1.2). All other combinations of metric, biodiversity measure and weighting scheme showed a pronounced influence of grain size on effect sizes, with underestimation the principal result except at large grains.

We show that full-data analyses are similarly sensitive to scale bias by way of shrinkage, unless properly specified with interaction terms that reflect scale-dependence of biodiversity change drivers. Full-data analyses that incorporated a forest-type-by-grain-size interaction performed similarly to meta-analyses of *LR* and appropriately revealed strong scale-dependence of forest type effects on species density, and its weak scale-dependence on asymptotic species richness estimates. In practice, however, model misspecification is likely if numerous covariates are under consideration. Many, if not all, covariates are likely to exert scale-dependent effects, but sample sizes will limit specification of all interactions and random slopes. We emphasise the importance of a principled causal investigation of focal treatment effects in full-data analyses. This is achieved at the stage of systematic review and critical appraisal, by rigorous selection of control and treatment groups that are matched with respect to other covariates that may exert scale-dependent effects.

### Best practice for biodiversity syntheses

Our simulations of meta-estimates and meta-regressions result in straightforward recommendations for analytical methods that minimise scale bias in meta-analysis of a treatment applied to a patch-level species pool. We advocate asymptotic measures of richness rather than species density, log response ratios rather than standardised mean differences, and random-effects weighting or no weighting rather than fixed-effects weighting.

For all cross-study syntheses of biodiversity differences, we recommend adherence to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), by reporting the risks of bias both within and across studies (Liberati *et al*. 2009). Cross-study scale biases are rarely acknowledged, yet can be revealed by testing for covariation among *A, N, V* and effect sizes. The estimate and CI of coefficients for regressions against plot size need interpreting with reference to the full suite of covariance amongst *A*, *N*, *V* and effect size, in order to evaluate the possibility of covariation obscuring scale dependence. We recommend that future meta-analyses publish these simple diagnostics to increase transparency about the potential for scale dependence to influence meta-estimation across heterogeneous studies. Similarly, full-data analysts can explore the degree of shrinkage of random, study-level intercepts and slopes in relation to *A*, *N* and *V*. Ultimately, the researcher needs to explore, describe and explain the potential for scale dependence, and then elaborate on its implications for management recommendations of relevance to decision-makers.

If scale dependence is detected using these diagnostics, subgroup-analysis may be required to understand how meta-estimates, and their response to covariates of interest, vary among subgroups of studies with similar grain. Although subgrouping will mitigate meta-estimation bias caused by the studies at the smallest scales, it will reduce sample sizes, and estimates based on *g* will not be directly comparable across groups, due to scale-dependence of replication or variance.

### Future work on scale dependence in biodiversity synthesis

Our simulations of tree communities in forest patches were deliberately simple in construction, to test for effects of study-level replication and grain size on cross-study estimation of effect size. Future simulations could investigate how variation in abundances, species abundance distributions (SADs) and aggregation influence the sensitivity of meta-estimates to choices made when conducting a synthesis. Our varying of only total treatment species richness limits the scope of inferences from our investigation to situations where total species richness differences are the predominant component of biodiversity change (Chase et al. 2019). Differences in other components will also drive differences in the magnitude and direction of scale-dependence and together might obscure the detection of scale-dependence across heterogeneous study pools (Chase et al. 2019).

We compared the performance of two commonly used biodiversity measures: species density and asymptotic richness estimated with the Chao-1 non-parametric estimator. Future research could examine alternative measures. For example, Chao & Jost (2012) advocate the use of a non-asymptotic standardization approach via coverage-based rarefaction and extrapolation, when available data do not contain sufficient information to accurately infer the true diversity of an entire assemblage.

The studies collated for meta-analysis can vary widely in design regarding their spatio-temporal interspersion of control and treatment units. Study designs that require randomisation (e.g. Randomised Controlled Trials) may be more likely to employ smaller plot sizes that can be readily randomised, while designs requiring greater temporal investment of effort (e.g. Before-After Control-Impact) may sample larger plots with fewer replicates. Christie et al. (2019) showed that simpler study designs can yield lower accuracy combined with lower variance, and thus a higher propensity for bias. We have shown that this bias will become amplified in precision-weighted meta-analyses. The combination of different levels of design-related study bias and the different grains at which studies implement their designs could lead to complex biases that may not balance out in meta-analyses, even when large numbers of studies are included. We therefore recommend using diagnostic plots of the sort that allowed us to explore potential design-dependent biases.

Empirical studies increasingly estimate richness rather than density, and now commonly archive their raw data, allowing meta-analysts to estimate richness from raw occurrences (Gerstner *et al.* 2017). These developments will help to reduce bias in meta-effect estimation in the future, if *PatchRich*E allows estimation of *LR*. Because meta-analyses typically collate studies spanning a broad time period, however, meta-analyses often include a mixture of species-richness metrics, with potential for more scale dependence in meta-effect sizes from pools of older studies using species density. This will likely be particularly problematic for research investigating temporal changes in effect sizes over time (Koricheva & Kulinskaya, 2019), making exploration of scale dependence critical in such instances.

## Concluding remarks

Open access to plot-level biodiversity data will likely increase the prevalence of full-data analyses with multilevel models. Meta-analysis will nevertheless remain a valuable tool, because it can synthesise studies that vary in response metric, and that report only summary statistics. Both meta-analyses and multilevel models treat the observed sample variance as a true variance, ignoring possible biases (Lin 2018). However, inaccurate variance estimates in little-replicated ecological studies are known to cause an accumulating bias in precision-weighted meta-analysis, requiring correction (Doncaster & Spake 2018). An equivalent correction awaits development for full-data analyses to under-shrink little-replicated studies. This may best suit Bayesian approaches, which can use priors to control how shrinkage is added to the model (Clark 2019).

## Acknowledgements

We thank S. Nakagawa, G. Stewart and W.J. Sutherland for insightful discussions. RS was funded by the Japan Society for the Promotion of Science (BRIDGE Fellowship), and the U.K. Biotechnology and Biological Sciences Research Council (BB/H531935/1). We thank the authors whose field data contributed to the two empirical meta-analyses.

## References

Antão, L.H., Bates, A.E., Blowes, S.A., Waldock, C., Supp, S.R., Magurran, A.E., *et al.* (2020). Temperature-related biodiversity change across temperate marine and terrestrial systems. *Nat. Ecol. Evol.*, 4, 927–933.

Avery, T.E. & Burkhart, H.E. (2015). *Forest Measurements. 5th edn.* Fifth. Waveland Press, Inc., Long Grove, Illinois.

Bates, D., Maechler, M., Bolker, B. & S, W. (2012). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, 67, 1–48.

Bell, A., Holman, D. & Jones, K. (2019). Using shrinkage in multilevel models to understand intersectionality: a simulation study and a guide for best practice. *Methodology*, 15, 88–96.

Blowes, S.A., Supp, S.R., Antão, L.H., Bates, A., Bruelheide, H., Winter, M., *et al.* (2019). The geography of biodiversity change in marine and terrestrial assemblages, *Science*, 345, 339–345.

Ceballos, G., Ehrlich, P.R. & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proc. Natl. Acad. Sci. U.S.A.*, 114, E6089–E6096.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.*, 265–270.

Chao, A. & Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93, 533–2547.

Chase, J.M. & Knight, T.M. (2013). Scale-dependent effect sizes of ecological drivers on biodiversity: Why standardised sampling is not enough. *Ecol. Lett.*, 16, 17–26.

Chase, J.M., Mcgill, B.J., Mcglinn, D.J., May, F., Blowes, S.A., Xiao, X., *et al.* (2018). Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. *Ecol. Lett.*, 21, 1737–1751.

Chase, J.M., Mcgill, B.J., Thompson, P.L., Antão, L.H., Bates, A.E., Blowes, S.A., *et al.* (2019). Species richness change across spatial scales. *Oikos*, 128, 1079–1091.

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I. & Sutherland, W.J. (2019). Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *J. Appl. Ecol.*, 56, 2742–2754.

Connell, J.H. & Sousa, W.P. (1983). On the Evidence Needed to Judge Ecological Stability or Persistence. *Ametrican Nat.*, 121, 789–824.

Curran, M., Hellweg, S. & Beck, J. (2014). Is there any empirical support for biodiversity offset policy? *Ecol. Appl.*, 24, 617–632.

Daskalova, G.N., Myers-smith, I.H., Bjorkman, A.D. & Blowes, S.A. (2020). Landscape-scale forest loss as a catalyst of population and biodiversity change. *Science (80-. ).*, 368, 1341–1347.

Doncaster, C.P. & Spake, R. (2018). Correction for bias in meta-analysis of little-replicated studies. *Methods Ecol. Evol.*, 634–644.

Duguid, M.C. & Ashton, M.S. (2013). A meta-analysis of the effect of forest management for timber on understory plant species diversity in temperate forests. *For. Ecol. Manage.*, 303, 81–90.

Field, R., Hawkins, B.A., Cornell, H. V., Currie, D.J., Diniz-Filho, J.A.F., Guégan, J.F., *et al.* (2009). Spatial species-richness gradients across scales: A meta-analysis. *J. Biogeogr.*, 36, 132–147.

Fletcher, D. & Dixon, P.M. (2012). Modelling data from different sites, times or studies: weighted vs. unweighted regression. *Methods Ecol. Evol.*, 3, 168–176.

Gardner, T.A., Barlow, J., Parry, L.W., Peres, C.A., Mpeg, G., Perimetral, A., *et al.* (2007). Predicting the Uncertain Future of Tropical Forest Species in a Data Vacuum, 39, 25–30.

Geary, R.C. (1930). The frequency distribution of the quotient of two normal variates. *J. R. Stat. Soc.*, 442–446.

Gerstner, K., Moreno-mateos, D., Gurevitch, J., Beckmann, M., Kambach, S., Jones, H.P., *et al.* (2017). 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.

Gibson, L., Lee, T., Koh, L., Brook, B., Gardner, T.A., J., B., *et al.* (2011). Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature*, 478, 378–381.

Goldfeld, K. (2019). simstudy: Simulation of Study Data. R package version 0.1.14.

Gonzalez, A., Cardinale, B.J., Allington, G.R.H., Byrnes, J., Endsley, K.A., Brown, D.G., *et al.* (2016). Estimating local biodiversity change : a critique of papers claiming no net loss of local diversity. *Ecology*, 97, 1949–1960.

Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, 4, 379–391.

Gotelli, N.J. & Colwell, R.K. (2011). Estimating species richness. In: *Biological Diversity Frontiers in Measurment and Assessment* (eds. Magurran, A.E. & Mcgill, B.J.). Oxford University Press, Oxford, pp. 39–54.

Grussu, G., Testolin, R., Saulei, S., Farcomeni, A., Yosi, C.K., De Sanctis, M., *et al.* (2016). Optimum plot and sample sizes for carbon stock and biodiversity estimation in the lowland tropical forests of Papua New Guinea. *Forestry*, 89, 150–158.

Guariguata, M.R., Chazdon, R.L., Denslow, J.S., Dupuy, J.M. & Anderson, L. (1997). Structure and floristics of secondary and old-growth forest stands in lowland Costa Rica. *Plant Ecol.*, 132, 107–120.

Gurevitch, J., Curtis, P.S. & Jones, M.H. (2001). Meta-analysis in ecology. *Adv. Ecol. Res.*, 32, 199–247.

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

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

Hamman, E., Pappalardo, P., Bence, J.R., Peacor, S.D. & Osenberg, C. (2018). Bias in meta-analyses using Hedges’ d. *Ecosphere*, 9, e02419.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychol. Bull.*, 92, 490–499.

Hedges, L. V. (1981). Distribution theory for Glass’s estimator of effect size and related estimators. *J. Educ. Stat.*, 6, 107–128.

Hedges, L. V. (1983). A Random Effects Model for Effect Sizes. *Psychol. Bull.*, 93, 388.

Hedges, L. V & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, San Diego, CA.

Hill, J.K. & Hamer, K.C. (2004). Determining impacts of habitat modification on diversity of tropical forest fauna : the importance of spatial scale, 744–754.

Hortal, J., Borges, P.A. V & Gaspar, C. (2006). Evaluating the performance of species richness estimators : sensitivity to sample grain size. *J. Anim. Ecol.*, 75, 274–287.

Ioannidis, J.P.A., Patsopoulos, N.A. & Rothstein, H.R. (2008). Reasons or excuses for avoiding meta-analysis in forest plots. *Br. Med. J.*, 336, 1413–1415.

IPBES. (2019). *The IPBES assessment report on land degradation and restoration.*

Jiménez, J.U., Fábrega, J., Mora, D., Tejedor, N. & Sánchez, M. (2016). Composition, diversity and tree structure of a tropical moist forest in Gamboa, Colon, Panama. *Air, Soil Water Res.*, 9, 29–24.

Jung, M., Rowhani, P. & Scharlemann, J.P.W. (2019). Impacts of past abrupt land change on local biodiversity globally. *Nat. Commun.*, 10, 5474.

Keil, P. & Chase, J. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nat. Ecol. Evol.*, 3, 390–399.

Koricheva, J., Gurevitch, J. & Mengersen, K. (2013). *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton University Press, Princeton and Oxford.

Koricheva, J. & Kulinskaya, E. (2019). Temporal instability of evidence base: a threat to policy making ? *Trends Ecol. Evol.*, 34, 895–902.

Král, K., Janík, D., Vrška, T., Adam, D., Hort, L., Unar, P., *et al.* (2010). Local variability of stand structural features in beech dominated natural forests of Central Europe: Implications for sampling. *For. Ecol. Manage.*, 260, 2196–2203.

Lajeunesse, M.J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, 96, 2056–2063.

Levin, S.A. (1992). The problem of pattern and scale in ecology. *Ecology*, 73, 1943–1967.

Levins, R. (1966). The strategy of model building in population biology. *Am. Sci.*, 54, 421–31.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., *et al.* (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions : explanation and elaboration. *J. Clin. Epidemiol.*, 62, e1–e34.

Lin, L. (2018). Bias caused by sampling error in meta- analysis with small sample sizes, 1–19.

Magurran, A.E. (2004). *Measuring biological diversity*. Blackwell, Oxford.

May, F., Gerstner, K., Mcglinn, D.J., Xiao, X. & Chase, J.M. (2018). mobsim : An r package for the simulation and measurement of biodiversity across spatial scales. *Methods Ecol. Evol.*, 9, 1401–1408.

McGlinn, D.J., Xiao, X., May, F., Gotelli, N.J., Engel, T., Blowes, S.A., *et al.* (2018). Measurement of Biodiversity (MoB): A method to separate the scale-dependent effects of species abundance distribution, density, and aggregation on diversity change. *Methods Ecol. Evol.*, 10, 58–269.

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

Michael Clark. (2019). *Shrinkage in Mixed Effects Models*. Available at: https://m-clark.github.io/posts/2019-05-14-shrinkage-in-mixed-models/. Last accessed .

Monserud, R.A. (2004). Large-scale management experiments in the moist maritime forests of the Pacific Northwest. *Landsc. Urban Plannning*, 59, 159–180.

Nakagawa, S., Dunn, A.G., Lagisz, M., Bannach-brown, A., Grames, E.M., Sánchez-tójar, A., *et al.* (2020). A new ecosystem for evidence synthesis.

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

Newbold, T., Hudson, L., Hill, S.L.L., Contu, S., Senior, R.A., Borger, L., *et al.* (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, 2, 45.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., *et al.* (2019). vegan: Community Ecology Package. R package version 2.5-5.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., *et al.* (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *Bioscience*, 51, 933.

Osenberg, C.W., Sarnelle, O. & Cooper, S.D. (1997). Effect size in ecological experiments: the application of biological models in meta-analysis. *Am. Nat.*, 150, 798–812.

Palma, A. De, Ortiz, K.S., Martin, P.A., Chadwick, A., Gilbert, G., Bates, A.E., *et al.* (2018). *Challenges With Inferring How Land-Use Affects Terrestrial Biodiversity : Study Design , Time , Space and Synthesis*. *Next Gener. Biomonitoring Part 1*. 1st edn. Elsevier Ltd.

Poole, C. & Greenland, S. (1999). Random-effects meta-analysis are not always conservative. *J. Epidemiol.*, 150, 469–475.

Powell, K.I., Chase, J.M. & Knight, T.M. (2011). A synthesis of plant invasion effects on biodiversity across spatial scales. *Am. J. Bot.*, 98, 539–548.

Powell, K.I., Chase, J.M. & Knight, T.M. (2013). Invasive plants have scale-dependent effects on diversity by altering species-area relationships. *Science (80-. ).*, 339, 316–318.

R Core Team. (2019). R: A Language and Environment for Statistical Computing. v. 3.5.3.

Ramage, B.S., Sheil, D., Salim, H.M.W., Fletcher, C., Mustafa, N.Z.A., Luruthusamay, J.C., *et al.* (2013). Pseudoreplication in tropical forests and the resulting effects on biodiversity conservation. *Conserv. Biol.*, 27, 364–372.

Rosenberg, M.S., Rothstein, H. & Gurevitch, J. (2013). Effect Sizes: Conventional Choices and Calculations. In: *Handbook of Meta-analysis in Ecology and Evolution*. Oxford, pp. 61–71.

Shimadzu H. (2018). On species richness and rarefaction: size- and coverage-based techniques quantify different characteristics of richness change in biodiversity. *J. Math. Biol.*, 77, 1363–1381.

Solt, F., Hu, Y. & Brenton Kenkel. (2018). interplot: Plot the Effects of Variables in Interaction Terms, v.0.2.2.

Spake, R., Bellamy, C., Graham, L., Watts, K., Wilson, T., Norton, L., *et al.* (2019a). An analytical framework for spatially targeted management of natural capital. *Nat. Sustain.*, 2, 90–97.

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

Spake, R., Ezard, T.H.G., Martin, P.A., Newton, A.C. & Doncaster, C.P. (2015). A meta-analysis of functional group responses to forest recovery outside of the tropics. *Conserv. Biol.*, 29, 1695–1703.

Spake, R., Yanou, S., Yamaura, Y., Kawamura, K., Kitayama, K. & Doncaster, C.P. (2019b). Meta-analysis of management effects on biodiversity in plantation and secondary forests of Japan. *Conserv. Sci. Pract.*, 1, e14; https://doi.org/10.1002/csp2.14.

Sreekar, R., Katabuchi, M., Nakamura, A., Corlett, R.T., Slik, J.W.F., Fletcher, C., *et al.* (2018). Spatial scale changes the relationship between beta diversity, species richness and latitude. *R. Soc. Open Sci.*, 5, 181168.

Sutherland, W.J., Pullin, A.S., Dolman, P.M. & Knight, T.M. (2004). The need for evidence-based conservation. *Trends Ecol. Evol.*

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.*, 36, 1–48.

Wiens, J.A. (1989). Spatial Scaling in Ecology. *Source Funct. Ecol.*, 3, 385–397.

## Figure legends

**Figure 1** Conceptual representation of typical study-design trade-offs. Data from individual field studies measuring forest management impacts on vegetation biodiversity (blue points, compiled by Duguid & Ashton, 2013). Here we index study-level ‘precision’ by replication, and ‘realism’ by plot size, both plotted on log scales. The meta-analysis has a replication-weighted average precision and realism across all *k* studies. This improves precision compared to the majority of studies, but reduces realism, because of the negative relationship of realism with precision in this instance, in common with many biodiversity syntheses. The synthesis increases ‘generality’, here indexed by number of terrestrial ecoregions (inset coloured regions, Olson *et al*. 2001). Generality can take various meanings for ecological contexts, including extent across taxa, biomes, space or time, and number of variables that modify effects (Gurevitch *et al*. 2018).

**Figure 2** Sources of scale dependence in meta-analyses of biodiversity differences

**Figure 3** Method of simulating tree communities within forest patches of two types: old-growth (control) and secondary (treatment) with different distributions of species richness. Schematic overview of the populations of interest (step 1), the studies of them (steps 2-4), and their cross-study synthesis (step 5). Here we assume equal-sized forest patches, such that a given value of *A* for grain size as a fraction of patch area corresponds to the same ground coverage in m2 across all patches in a study.

**Figure 4** Relationship of study-level accuracy to precision depends on purpose and method of effect-size estimation, for studies with sampling grain *A* and replication *N*. For purposes of meta-analysis, plots show the relationship using a) Hedges’ *g* with *Vg*; b) Hedges’ *g* with *Vg*\_alt; c) log response ratio *LR* with *VLR*. For purposes of full-data analysis, plots show the relationship using d) study means of plot-level biodiversity in old-growth and secondary forests (note different accuracy range to a-c). Left panels of each pair of plots show biodiversity measured as species density (*PlotDensity*E); right panels show species richness (*PatchRich*E).

**Figure 5** Accuracy of cross-study estimation depends on type and method of synthesis. Accuracies are plotted as a function of the *A*-by-*N* correlation coefficient ρ. For meta-effects from meta-analysis (a-c), accuracy depends on biodiversity measure (columns), metric (colours), weighting (rows) and *N*-by-*A* correlation (slopes). For population means from full-data analysis (d), accuracy depends on biodiversity measure only. Non-parametric smooth curves were computed by locally estimated scatterplot smoothing.

**Figure 6** The influence of sampling grain size on effect size depends on the type and method of synthesis. For random-effects meta-analysis (a-c), meta-regression slopes depend on metric (rows) and measure (columns). For full-data analysis, the effect of forest type, conditional on grain size, depend on the biodiversity measure. Horizontal dashed lines show true values set by µ*i* (as defined in the Methods).

**Figure 7** Influence of grain on effect size depends on effect-size metric (columns) and weighting scheme (rows), for forest management effects on plant species density. Meta-regressions have 95% prediction intervals (grey shading) based on uncertainty only in the plot-size effect. Point size is proportional to relative study weight for each meta-regression, with colours distinguishing different publications. Inset numbers on each plot give the regression coefficient (*β*) and significance of the plot-size effect (*P* < 0.05 in bold). Variances for *g* were estimated by the conventional *Vg*.

**Figure 8** Schematic overview of four steps to minimising and diagnosing scale-dependence in cross-study syntheses of biodiversity differences.

## Tables

Table 1. Glossary of terms used in the simulation

|  |  |  |
| --- | --- | --- |
| **Term** | **Notation** | **Definition** |
| **Biodiversity measures** |
| Species density | *PlotDensity*E | The observed number of species contained within a plot sampled from a patch |
| Estimated species richness | *PatchRich*E | Estimated minimum number of species in a sampled patch, applying the Chao-1 richness estimator to plot-level data |
| True species richness | *PatchRich*T | True number of species in a forest patch, set by the simulation; only known empirically from complete sampling of the patch |
| **Meta-analysis metrics** |
| Hedges’ *g* | *g* | *J·d*, where *d* is the standardised mean difference(‾*x*1– ‾*x*2) / *spooled* in which‾*x*1and‾*x*2are the biodiversity means of the treatment and control groups respectively, *spooled* is the pooled standard deviation of treatment and control groups given by weighting the standard deviations of each group by their sample sizes, *N*1 and *N*2, and *J* is a correction for low within-group replication, which takes the value:1 – 3/[4(*N*1 + *N*2 – 2) – 1]*.* |
| Variance in Hedges’ *g* | *Vg* | *J* 2 ([*N*1 + *N*2] / [*N*1*N*2] + *d*2 / (2[*N*1 + *N*2])) |
| Alternative variance in Hedges’ *g* | *Vg\_*alt | *J* 2 ([*N*1 + *N*2 – 2]/ (*N*1*N*2 [*N*1 + *N*2 – 4] / [*N*1 + *N*2])) |
| Log response ratio | *LR* | ln(‾*x*1/‾*x*2) |
| Variance in log response ratio | *VLR* | *s*2*pooled* (1 / [*N*1‾*x*12] + 1 / [*N*2‾*x*22]) |
| True population effect size  | *g*T,*LR*T | Effect size (*g* or *LR*) calculated from the population ,  and *N* of *PatchRich*T for 100 old-growth and secondary forest patches, set by the simulation |
| Estimation accuracy | (*g* – *g*T)/*g*T,(*LR* – *LR*T)/*LR*T | The standardized difference of the estimated from the true effect size, with a small value signifying high accuracy, and a positive (negative) value over- (under-) estimation. |
| Estimation precision  | 1/ *Vg*,1/ *Vg\_alt*,1/*VLR* | The spread of replicate effect-size estimates around their mean, with a large value signifying high precision.  |
| **Full-data analyses metrics** |
| Response |  | Biodiversity measured as *PlotDensity*E or *PatchRich*E |
| Variance |  | The average of the squared differences from the mean biodiversity value, among treatment replicates within a study (as used in Fig. 4d) |
| Random groups |  | Categorical variable, with levels corresponding to the identities of the component studies |
| Population mean effect size |  | The cross-study estimate of biodiversity difference between treatment and control forest types (illustrated in Fig. 3.5b) |
| Conditional effect |  | The estimated, fixed effect of forest type, conditional on sampling grain. Estimated from full-data analyses that specified an interaction between forest type and grain size. |
| Estimation accuracy |  | The accuracy of the estimated, fixed population mean effect of forest type (the forest type coefficient). For models of *PlotDensity*E, the accuracy of forest type effects were estimated as for *LR* (above). For models of *PatchRich*E, the accuracy of forest type effect was standardised against the true species richness difference of 50 species. |