Understanding 'it depends' in ecology: a guide to hypothesising, visualising and interpreting statistical interactions

Spake, R. ORCID: https://orcid.org/0000-0003-4671-2225, Bowler, D. E., Callaghan, C. T., Blowes, S. A., Doncaster, C. P., Antao, L. H., Nakagawa, S., McElreath, R. and Chase, J. M. (2023) Understanding 'it depends' in ecology: a guide to hypothesising, visualising and interpreting statistical interactions. Biological Reviews, 98 (4). pp. 983-1002. ISSN 1469-185X doi: https://doi.org/10.1111/brv.12939 Available at https://centaur.reading.ac.uk/110540/

BIOLOGICAL REVIEWS
Cambridge Philosophical Society
Biol. Rev. (2023), pp. 000–000.
doi: 10.1111/brv.12939

Understanding 'it depends' in ecology: a guide to hypothesising, visualising and interpreting statistical interactions

Rebecca Spake[1,2,*], Diana E. Bowler[1,3], Corey T. Callaghan[1,4,5], Shane A. Blowes[1,6], C. Patrick Doncaster[7], Laura H. Antão[8], Shinichi Nakagawa[9], Richard McElreath[1,10] and Jonathan M. Chase[1,6]

[1]*German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103, Leipzig, Germany*
[2]*School of Biological Sciences, University of Reading, RG6 6EX, Reading, UK*
[3]*UK Centre for Ecology & Hydrology, OX10 8BB, Oxfordshire, UK*
[4]*Institute of Biology, Martin Luther University Halle – Wittenberg, 06120, Halle (Saale), Germany*
[5]*Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, University of Florida, Davie, 33314-7719, FL, USA*
[6]*Department of Computer Science, Martin Luther University Halle-Wittenberg, 06099, Halle (Saale), Germany*
[7]*School of Biological Sciences, University of Southampton, SO17 1BJ, Southampton, UK*
[8]*Research Centre for Ecological Change, Faculty of Biological and Environmental Sciences, University of Helsinki, 00014, Helsinki, Finland*
[9]*UNSW Data Science Hub, Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, UNSW, Sydney, 2052, NSW, Australia*
[10]*Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig, 04103, Germany*

ABSTRACT

Ecologists routinely use statistical models to detect and explain interactions among ecological drivers, with a goal to evaluate whether an effect of interest changes in sign or magnitude in different contexts. Two fundamental properties of interactions are often overlooked during the process of hypothesising, visualising and interpreting interactions between drivers: the measurement scale – whether a response is analysed on an additive or multiplicative scale, such as a ratio or logarithmic scale; and the symmetry – whether dependencies are considered in both directions. Overlooking these properties can lead to one or more of three inferential errors: misinterpretation of (*i*) the detection and magnitude (Type-D error), and (*ii*) the sign of effect modification (Type-S error); and (*iii*) misidentification of the underlying processes (Type-A error). We illustrate each of these errors with a broad range of ecological questions applied to empirical and simulated data sets. We demonstrate how meta-analysis, a widely used approach that seeks explicitly to characterise context dependence, is especially prone to all three errors. Based on these insights, we propose guidelines to improve hypothesis generation, testing, visualisation and interpretation of interactions in ecology.

Key words: antagonistic, effect size, generalised linear models; Hedges' *g*, log response ratio, meta-regression, statistical interaction, synergistic, synthesis, transformation.

CONTENTS

* Author for correspondence (E-mail: r.spake@reading.ac.uk).

I. INTRODUCTION

Nature's complexity and multi-causality frequently leads ecologists to describe ecological relationships as being 'context dependent' (e.g. Spake *et al.*, 2022*a*,*b*; Bradley *et al.*, 2020; Catford *et al.*, 2021; Wirsing *et al.*, 2021). The term 'context dependence', derived from the Latin words *con* ('together'), *texere* ('to weave'), and *pendere* ('to hang'), aptly describes the central remit of ecology: to characterise relationships among the causal threads of life's rich tapestry. Ecologists might ask, for example, how do landscape attributes modify the impact of organic farming on biodiversity (Seufert & Ramankutty, 2017; Smith *et al.*, 2020)? How are biodiversity effects on ecosystem functioning modified by global-change drivers such as drought (Hong *et al.*, 2022)? How does the impact of invasive species on native biodiversity depend on the spatial grain at which it was measured (Powell, Chase & Knight, 2011)? Studying the dependencies among ecological drivers has both practical and theoretical motivations. For example, identifying interacting effects can help target limited conservation resources to contexts where interventions will be most effective (Spake *et al.*, 2019), while the absence of interactions might suggest the existence of general relationships in ecology (Leimu *et al.*, 2006).

In this review, we address a common approach to the investigation of context dependence, which asks whether a driver of interest has an effect that 'depends on', or gets 'modified' in magnitude or sign by, other drivers (Vanderweele, 2009, 2019). In principle, this line of questioning might seem straightforward and amenable to statistical testing with ecological data, by fitting models containing 'statistical interactions' (see Table 1 for glossary), or by using meta-analytic methods that explore whether 'effect sizes' systematically vary across putative ecological gradients or factors (Gurevitch *et al.*, 2018; Spake *et al.*, 2022*b*). Such analyses are vulnerable to several potential misinterpretations, however, which arise when two critical aspects of effect modification are overlooked: *scale* (whether a response is analysed on an additive or multiplicative scale) and *symmetry* (whether effect modification is examined in both directions).

The nature of effect modification depends on the measurement scale used in the analysis – that is, whether a response is analysed on an additive scale or a multiplicative scale, such as a ratio, logarithmic or logit scale (Vanderweele, 2009; Greenland, 2015). Ecological data often do not conform to the assumptions of linear models, requiring the use of transformations (e.g. log-transformation) or non-linear link functions (Bolker *et al.*, 2009). Such transformations, however, can change the functional form of the relationships between response and predictor variables, and influence qualitative and quantitative inferences about effect modification (Wagenmakers *et al.*, 2012; VanderWeele & Knol, 2014). This change is often not explicitly considered during interpretation (Griffen *et al.*, 2016), which can lead to erroneous inferences about the detection and magnitude (henceforth 'Type-D' errors, where D denotes detectability issues) and

Table 1. Glossary.

BACI design	Before–After–Control–Intervention. The outcomes of an intervention treatment are compared to those in a control, with both referenced against pre-treatment responses to account for unmeasured environmental variation. In these designs, the effect of an intervention is measured not in its main effect (CI), but in its interaction with time (the BA × CI term).
Conditional plots	Predicted values of Y plotted across the range of X and Z, or at substantively meaningful values of these predictors.
Effect sizes (in meta-analyses)	Effect sizes estimate the magnitude and direction of change in a response variable Y, either as differences between categorical group means or as the strength of relationships for a continuous focal driver.
Marginal effect	Marginal effects summarise the effect of an independent variable on the response in terms of a model's predictions. Marginal effect plots display the estimated coefficient of a focal variable and its confidence interval against values of a modifying variable. They indicate the statistical significance, uncertainty, magnitude, and direction of an effect across a full hypothetical range of the modifying variable, often a range from 3SD below to 3SD above the mean. Best practice is to include a frequency histogram of the modifying variable along the x-axis, to allow the user to judge common support based on the distribution of the modifier.
Scale of measurement	The scale on which an effect is estimated, generally either additive or multiplicative.
Statistical interaction	A statistical interaction involves the effect of each explanatory variable on the response varying with the magnitude or sign of other variables. The magnitude and sign of interaction can depend on the scale of measurement (whether multiplicative or additive). Detection of a statistical interaction does not necessarily imply a biological interaction, for example if the interaction is enforced by ceiling/floor constraints on the response variable.

sign ('Type-S' errors, where S denotes sign issues) of effect modification (Fig. 1). In ecology, D and S errors are typically discussed in relation to modelling biases that arise from measurement error and low statistical power (Duncan & Kefford, 2021; Yang *et al.*, 2022), but they can also stem from model misinterpretation (Duncan & Kefford, 2021; Wolkovich *et al.*, 2021). For example, the choice of measurement scale has affected the interpretation of temporal trends in biodiversity indices (e.g. Leung *et al.*, 2020; Loreau *et al.*, 2022), and temperature sensitivities of organisms to warming (Wolkovich *et al.*, 2021). As an illustration, consider temporal trends in species richness at a location for two taxonomic groups. Richness in group A might decline by 30%, while group B might decline by 50%. If group A is considerably more speciose than B, then its smaller percentage decline may nevertheless correspond to a greater absolute loss of species. If the analyst is interested in relating local extinction rates over time to predictors such as group-level traits, D and S errors can result from interpreting such losses as percentages only.

Ecologists often approach interactions asymmetrically, to construct hypotheses about a 'focal' driver of interest (X) and its modification by a second, 'modifying' variable (Z) that is often beyond the control of the researcher (Cox, 1984). This asymmetry of focus often leads the analyst to generate hypotheses and predictions about effect modification in a single direction (Berry, Golder & Milton, 2012). As an example, one might ask how biodiversity effects on ecosystem functioning are modified by environmental stress. Statistically, however, effect modification is symmetric: if Z modifies the effect of X, then X modifies the effect of Z. If the effect of biodiversity on ecosystem functioning depends on the level of environmental stress, then the effect of environmental stress on functioning depends on the level of biodiversity. Thus, interpreting and visualising dependencies in a single direction may be insufficient for testing hypotheses, when it overlooks patterns that are inconsistent with the underlying conditional theory (henceforth 'Type-A' errors, where A denotes asymmetry issues; Berry *et al.*, 2012; Fig. 1). Asymmetric approaches to effect modification are inherent to the method of meta-analysis, which estimates the magnitude of focal effects across individual studies (effect sizes), and evaluates their variation with putative 'effect modifiers'. Meta-analysis is a widely used approach in ecology (Gurevitch *et al.*, 2018; Anderson *et al.*, 2021), which often explicitly sets out to test for and explain context dependence in ecological effects (e.g. Leal & Peixoto, 2017; Marino, Romero & Farjalla, 2018; Albertson *et al.*, 2021). The consequences of asymmetric investigation for ecological inference have yet to be evaluated.

Here, we review the inferential errors that can arise when the scale and symmetry of effect modification are overlooked in ecological studies. Several statistical challenges to modelling context dependence in ecology have previously been recognised in relation to confounding variation, collinearity and statistical power (e.g. Catford *et al.*, 2021; Duncan & Kefford, 2021). We extend the list of challenges to Type-D, -S and -A errors, and provide widely applicable principles and practical guidance for improving the study of interactions across a variety of ecological questions. We begin by illustrating with empirical data and simulations how D, S and A errors can result from ignoring the scale and symmetry of interactions, even when a model is properly specified. We then demonstrate how meta-analysis is particularly vulnerable to these errors despite its wide use and often explicit goal to evaluate and understand context dependence. Based on these insights, we outline key considerations to improve hypothesis generation and testing, as well as the visualisation and interpretation of conditional effects in ecology.

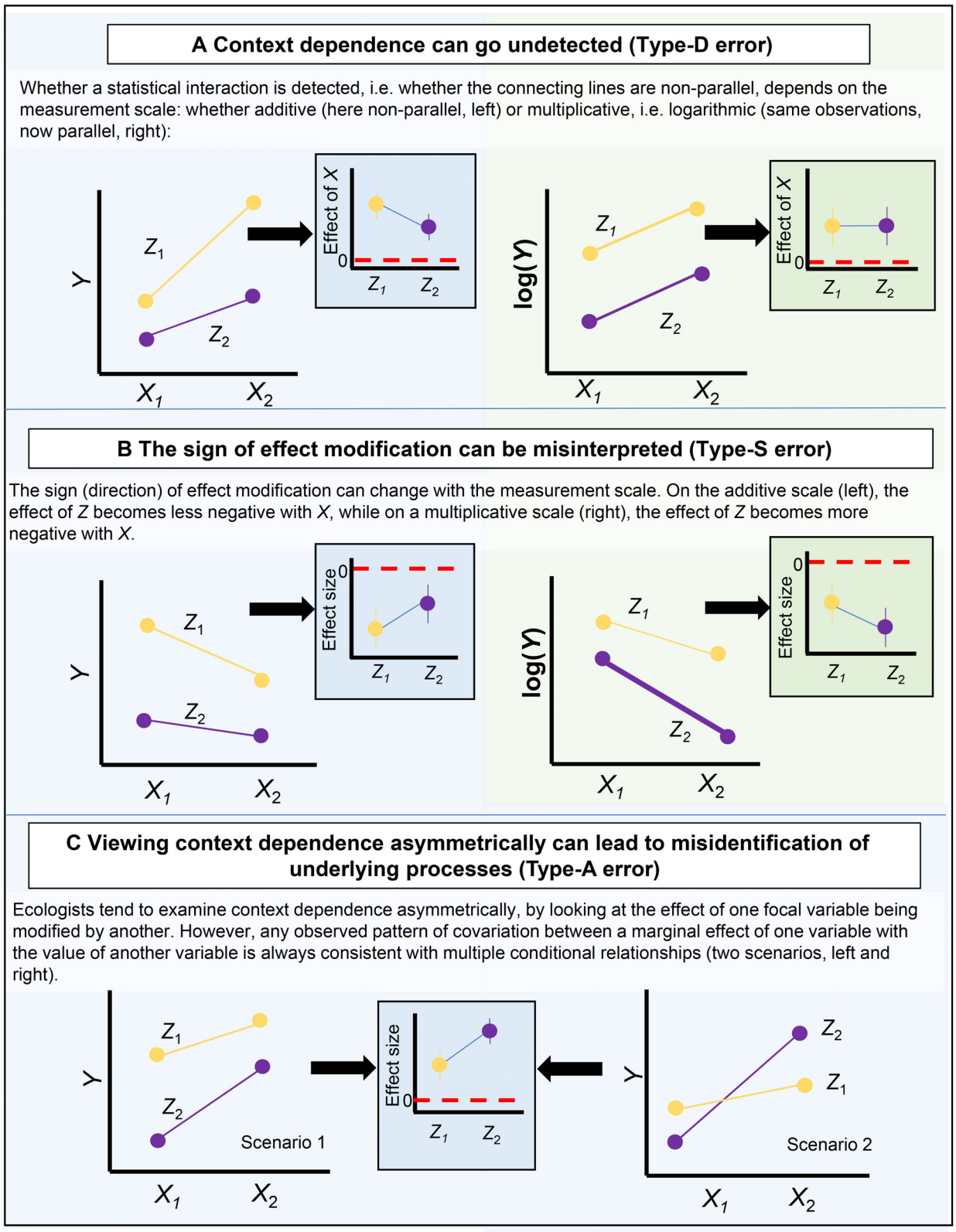

Fig. 1. Three common inferential errors when investigating context dependence in ecology. Consider a test of context dependence in its most basic form: a 2 × 2 factorial experiment, measuring an ecological response Y, to the crossing of factors X and Z, each with two levels. The analyst fits a statistical model with an interaction term to the data: $Y \sim X + Z + X \times Z$, to test for and quantify context dependence. Three inferential errors are possible when the measurement scale or symmetry of the interaction are overlooked: detection and magnitude (Type D), sign (Type S) and misidentification of underlying processes (Type A).

II. CONTEXT DEPENDENCE CAN GO UNDETECTED (TYPE-D ERROR)

The most common way to test for context dependence is by introducing a statistical interaction ($X \times Z$) into a model. Statistical interactions indicate that the relationship between X and Y varies throughout the range of Z; and likewise, Z–Y relationships vary across the range of X (Duncan & Kefford, 2021). For example, if the effect of biodiversity on ecosystem functioning depends on the level of environmental stress, than the effect of environmental stress on functioning depends on the level of biodiversity. The statistical support for an interaction is then determined by evaluating its statistical significance (e.g. $P < 0.05$), or using model selection criteria to justify its inclusion in competing models (e.g. Akaike's information criterion, AIC). However, whether or not an interaction term is supported can depend critically on the measurement scale used to estimate the effects in a statistical model, i.e. whether the measurement scale is additive (e.g. absolute units) or multiplicative (e.g. log-transformed).

(1) Statistical interactions are scale dependent

To demonstrate this type of scale dependence, consider a statistical interaction in its most basic form: a 2 × 2 factorial experiment, measuring an ecological response Y, to the crossing of factors X and Z. An interaction is detected (and the null hypothesis rejected) when the lines on an interaction plot (connecting same-level means of one factor across levels of the other) are not parallel, even after accounting for uncertainties in the sizes of mean values (Fig. 1; Wagenmakers *et al.*, 2012). In this case, the effects of each factor differ according to the level of the other factor. However, the degree of parallelism can depend on the measurement scale (Figs 1B and 2). Additive scales measure change in equal increments along the range of a variable (e.g. biomass change in grams), whereas multiplicative scales measure relative change (e.g. per cent change in biomass relative to a control or baseline value). For purely mathematical reasons, if both X and Z affect Y independently, an absence of effect modification of the absolute difference with Z (i.e. parallel lines on an additive scale) forces relative measures of the effect of X on Y to vary with Z (i.e. non-parallel on a multiplicative scale), and *vice versa* (Vanderweele, 2009; VanderWeele & Knol, 2014).

Not all statistical interactions are equally vulnerable to non-detection. To identify situations where important contingencies may go undetected, Loftus (1978) distinguished between 'non-removable' and 'removable' interactions (Fig. 2). A non-removable interaction involves a change in the sign of an effect, and can never be undone by an arbitrary smooth monotonic transformation, and is therefore also known as 'crossover' or 'qualitative' (Wagenmakers *et al.*, 2012; VanderWeele & Knol, 2014). As an example of a non-removable interaction, the effect of canopy cover on forest susceptibility to bamboo invasion (measured as a probability) is negative in warm regions of Japan, but positive in cool areas, where bamboo exhibits photoinhibition and its establishment is facilitated by denser forest canopies (Spake *et al.*, 2021*b*). The change in sign is unaltered by transformations of forest susceptibility. By contrast, a non-crossover or 'removable' interaction can be undone by a transformation of the measurement scale (Fig. 2). It is the removable interactions that are particularly vulnerable to Type D (and S) errors. This is because ecologists often ignore the measurement scale when interpreting fitted models, and exclude statistical interactions if they fail a significance test (e.g. $P < 0.05$), or use model selection criteria that employs penalties to compensate for the over-fitting of more complex models (e.g. AIC). We thus focus our review on, and give examples of, removable interactions in the following sections.

(2) The modelling scale can change the meaning of a statistical interaction

Ecologists often ignore the measurement scale when interpreting interactions, describing effects as 'stronger' or 'weaker' in different contexts. The modelling scale is often chosen to satisfy modelling assumptions or to improve model fit, yet it fundamentally changes the underlying form of the model fitted (Spake *et al.*, 2022*a*), as well as the meaning of the statistical interaction tested (Rothman, 2002). Interaction on an additive scale (i.e. absolute units) means that the combined effect of two predictors is larger (or smaller) than the sum of the individual effects of the two predictors, whereas interaction on a multiplicative scale (e.g. log-transformed) means that the combined effect is larger (or smaller) than the product of the individual effects. As a result, the meaning of statistical interaction terms varies between linear and generalised linear models (and among different link functions), which are frequently used in ecology.

Linear models with interaction terms take the following form:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) \quad (1)$$

The $X \times Z$ term allows both the intercept and the effect (slope) of X on $E[Y|X]$ to vary with different levels of Z. β terms refer to parameters to be estimated. Its statistical significance indicates that the combined effect of X and Z is larger (or smaller) than the sum of their individual effects.

Ecological variables typically respond non-linearly to environmental gradients and can be subject to ceiling and floor effects for those that are naturally bounded (e.g. survival rates bounded between 0 and 100%, or abundances bounded to be positive). Because of this bounding, ecologists often use a multiplicative scale for statistical analysis, for example by transformation of the response variable, or fitting generalised linear models with non-linear link functions.

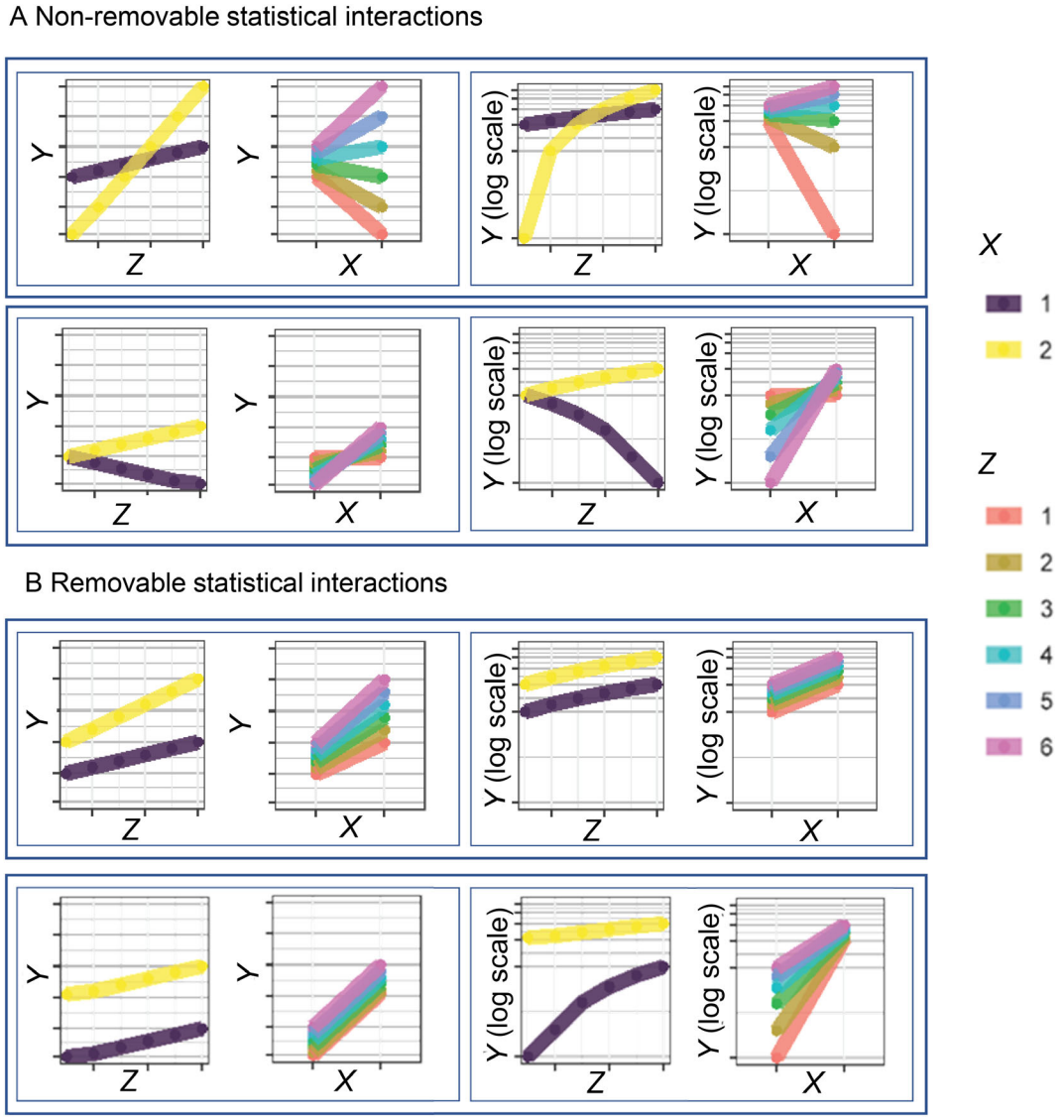

Fig. 2. Examples of 'removable' and 'non-removable' statistical interactions. These interaction plots display covariation of Y for two factor levels of X (distinguished by purple and yellow lines), and covariation in Y for six levels of Z (distinguished by multiple coloured lines). Each row corresponds to a separate example. On the left, Y is plotted on an additive scale, while the right panels display Y on a multiplicative, log scale. Non-removable interactions (A) cannot be undone by a transformation of the measurement scale, while removable interactions can (B).

In a generalised linear model (GLM) with a general functional form $f(.)$, the conditional expected value of Y (i.e. Y, given some value of Z and X) takes the following form:

$$E[Y \mid Z, X] = f(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z)) \quad (2)$$

where f could be any non-linear function, such as inverse-logit or inverse-logarithmic (exponential). In contrast to a linear model, the marginal effect of a predictor variable in a GLM is not constant over its range, or the range of other covariates (Karaca-Mandic, Norton & Dowd, 2012; Mize, 2019). Consider a binary logistic model with a response variable Y representing the conditional probability that a given binary outcome Y is equal to 1, $\Pr(Y = 1)$ (e.g. species presence), as a function of a continuous predictor X (e.g. an environmental gradient), and a categorical predictor Z (e.g. functional group), containing no interaction terms:

$$\Pr[Y \mid Z, X] = f(\beta_0 + \beta_1 X + \beta_2 Z) \quad (3)$$

The functional relationship $[f(.)]$ between X and $\Pr(Y = 1)$ is S-shaped for both levels of Z (Fig. 3). This means that for both levels of Z, an additional unit of X (i.e. the marginal effect of X) has little effect on $\Pr(Y = 1)$ for extremely high

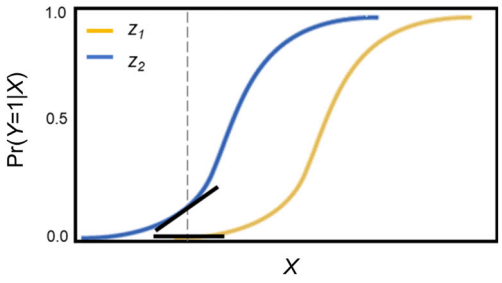

Fig. 3. Predicted probabilities of Y ($\Pr(Y = 1)$) obtained from a logistic regression model with continuous X and Z predictors and no statistical interaction. The marginal effect of X is shown at the vertical dashed line by the slope of the tangent to the S-shaped curve.

and low values of X, while the marginal effect of X is larger for intermediate values. The marginal effect of X is shown by the slope of the tangent to the S-shaped curve at a given value of X (e.g. at the vertical dashed line in Fig. 3). Despite there being no interaction term specified in the model, the marginal effect of X depends on the level of Z.

The existence of the non-linear link function in GLMs means that the effect of any predictor on the conditional expected value of Y depends on the values of every other explanatory variable (Berry *et al.*, 2012). In other words, GLMs with non-linear link functions, which include the canonical choices for Poisson (log link) and binomial (logit link) distributions, are inherently interactive in all of the predictors, even without interaction terms (Karaca-Mandic *et al.*, 2012). This changes the meaning of the interaction term: interaction on a multiplicative scale means that the combined effect is larger (or smaller) than the *product* of the individual effects (Rothman, Greenland & Walker, 1980; Knol *et al.*, 2007). It follows that hypotheses about contingent effects in non-linear systems, where GLMs are used, should specify expected marginal effects at particular values or distributions of all predictor variables (McCabe *et al.*, 2022).

Epidemiologists have long discussed the importance of scale for detecting and interpreting interactions (Rothman *et al.*, 1980; VanderWeele & Knol, 2014). The detection of interactions between binary risk factors (e.g. smoking status and asbestos exposure) for a binary outcome such as a mortality can depend on whether multiplicative, ratio measures (relative risks, risk ratios, and rate ratios), or additive difference measures (risk and rate differences) are used (Spiegelman & VanderWeele, 2017). Binomial models (with the logit link function), most often used for binary outcomes, implicitly measure interaction on the multiplicative scale (Fig. 3), yet additive scales that estimate the risk or rate differences (e.g. in years of life lost), are considered more policy relevant. Epidemiological guidelines consequently recommend presenting interaction analyses in a way that allows readers to assess interaction measures on multiple scales, and to assess additive interaction from multiplicative models (Knol *et al.*, 2011; Knol & VanderWeele, 2012).

(3) Is the additive or the multiplicative measurement scale more meaningful?

The inherent scale dependence of effect modification raises the question: on which scale should we interpret context dependence? The importance of distinguishing between the scale of interest and scale of measurement is both well recognised and much debated in epidemiology (Knol *et al.*, 2011). Many advocate the additive scale as the most policy-relevant (Hallqvist, Ahlbom & Reuterwall, 1996; VanderWeele & Robins, 2007), for targeting subgroups to maximise public health impact when resources are constrained (Knol *et al.*, 2011; Vanderweele, 2019). For example, if a public health study sets out to quantify how many lives might be saved by a policy intervention in different contexts, the absolute change in deaths on the additive scale will be of interest. The view of many epidemiologists is that it is almost always best to present both additive and multiplicative measures of interaction (Knol & VanderWeele, 2012). Similarly, for ecological questions, both scales are also likely to be informative for interpretation. For example, for biodiversity variables such as species richness and abundance, additive scales inform on changes in the absolute numbers of species or individuals, which may be of most interest when deciding between alternative local conservation actions, whilst multiplicative scales tell us about processes such as rates of population growth, which might be of most interest when examining drivers of population dynamics. Statistical significance testing requires meeting model assumptions, which may impose a measurement scale different to the interpretation scale. Having detected an effect, its biological meaning might be interpreted on only one or both scales, depending on the question. Ultimately, we must not conclude anything about scientific or practical (in)significance based on statistical (in)significance alone (Wasserstein, Schirm & Lazar, 2019; Abadie, 2020), and should aim to avoid overinterpretation (Mayo & Hand, 2022).

Transformations of the measurement scale can have important practical implications. For applied ecological questions, the consequences of failing to detect effect modification(s) could be more harmful than falsely detecting them, due to the ecological and economical costs of failing to take action or to better target them. For example, concluding that the effect of conservation intervention X on the establishment probability Y of a rare species is consistent across land-use intensity gradient Z (by way of a non-significant statistical interaction) could lead to missed opportunities to target conservation resources to sites with the greatest potential for conservation to enhance the likelihood of establishment. Similarly, when analysing data obtained from Before–After–Control–Impact (BACI) designs that are commonly employed in conservation research (reviewed by Wauchope *et al.*, 2021), the statistical significance of the interaction term is used to evaluate the effect of a conservation action. In these designs, the effect of an intervention is measured not in its main effect (C–I), but in its interaction with time (the BA × CI interaction term; Smokorowski & Randall, 2017).

Concluding that there was no effect of an intervention based on a statistical model employing a multiplicative scale might lead to missed opportunities to enhance absolute numbers of individuals or species, if such an effect was present, yet missed, on the additive scale.

Here, we advocate that under many circumstances, it is worthwhile to consider both additive and multiplicative scales. The following two examples illustrate why both measurement scales can be informative for assessing context dependence in ecology.

(*a*) *Empirical example: spider catch variation with artificial light at night and time of day*

In a factorial study that measured invertebrate abundance responses to time of day and to artificial-light exposure at night (Manfrin *et al.*, 2017), the abundance of a night-active ground-dwelling spider (*Pachygnatha clercki*) increased with the night-time artificial-light treatment (Fig. 4A). On the additive scale, in terms of the absolute number of spiders per unit effort, the same increase was observed in samples collected during both day and night. By contrast, on the multiplicative scale, the relative increase was greater between samples collected during the day, when catches were generally lower for this night-active species. Absolute and relative effects were different because of differences in mean spider abundances for each factor level. We might expect to see density changes on a multiplicative scale if the changes are brought about by population growth or spider activity, but on the additive scale if changes result from external immigration of individuals. Thus, if measuring long-term effects of artificial light at night on a closed population, we might want to interpret the multiplicative scale that accounts for the bounded and non-linear nature of population growth; even if focused on short-term effects, we might hypothesise that light affects the activity patterns of individuals on a *per capita* basis, and therefore still use a multiplicative scale (Fig. 4A). However, if abundance is considered a proxy for how individuals redistribute themselves in response to light, then we would want to interpret the additive scale.

(*b*) *Empirical example: ant species richness variation with land-use intensity and exotic ground cover*

In a study sampling ant communities across gradients of land-use intensity and exotic ground cover, a generalised linear model fitted with a logarithmic link function detected no interaction between exotic cover and land-use intensity in their effects on ant species richness (Oliver *et al.*, 2016). Indeed, in a conditional effect plot that displays the predicted species richness against exotic cover at high and low levels of land-use intensity, the lines are parallel on the multiplicative (logarithmic) scale (Fig. 4B). This means, however, that the lines must diverge on the additive scale, where the absolute increase in species richness is greater in non-intensive land uses, even as the relative change is roughly constant. In this case, given that the treatments have similar species richness, the difference between the predictions on each scale do not greatly differ.

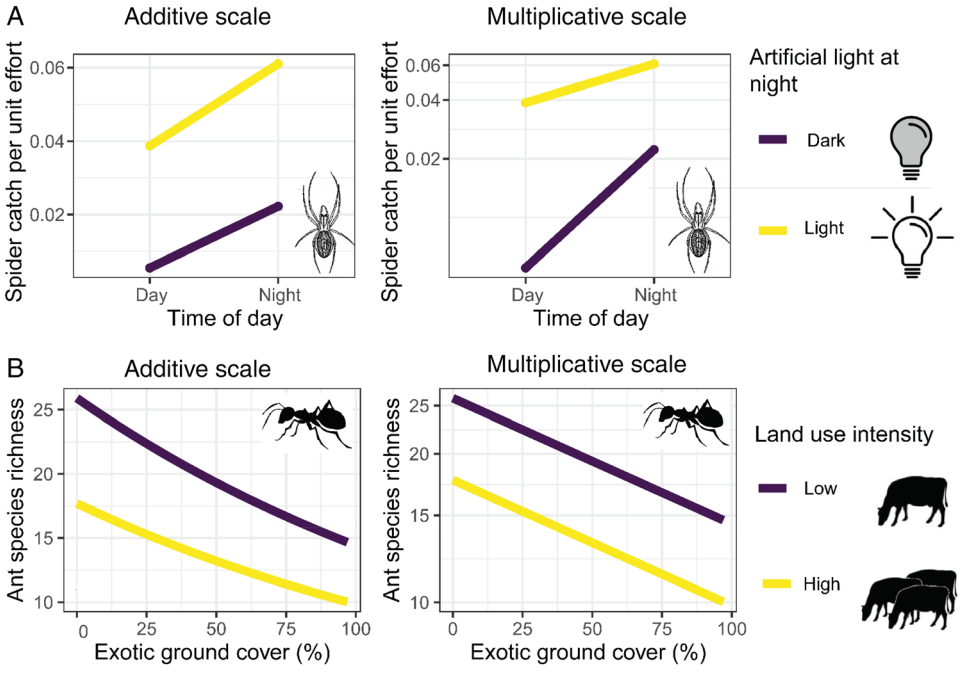

Fig. 4. The detection of interactions can depend on the measurement scale. Contingency is detected when the lines representing different treatment levels (purple *versus* yellow) are not parallel (i.e. statistically different slopes). (A) Spider catch per unit effort at day and night, shown for a site exposed to artificial light at night (yellow) and a dark control site (purple). Data from Manfrin *et al.* (2017). (B) Ant richness variation with exotic ground cover shown for high and low land-use intensity. Data from Oliver *et al.* (2016). See Section III for interpretation.

Nonetheless, interpretation on the additive scale might be more informative if a conservationist aims to target interventions to land uses that yield the greatest species richness increase, in which case reductions in exotic ground cover will yield the greatest absolute increase in low-intensity land uses.

III. THE DIRECTION OF EFFECT MODIFICATION IS VULNERABLE TO MISINTERPRETATION (TYPE-S ERROR)

(1) The sign of effect modification is scale dependent

In addition to the detection and magnitude of effect modification, its sign can also depend on the measurement scale. In other words, whether we conclude that the effect of X gets smaller or larger with Z can depend on whether we use an additive or multiplicative scale. This change in sign tends to occur when ecological response variables span orders of magnitude.

(*a*) *Empirical example: moth species richness over time across Finland*

We re-analysed the species richness data of moths published in Leinonen *et al.* (2016) and Antão *et al.* (2020), spanning 17 years across a latitudinal gradient in Finland. Following typical practice, we fitted a generalised linear multilevel model with a logarithmic link function to the Poisson-distributed counts of species, specifying an interaction between latitude and sampling year, and specifying that each trap has a varying (random) intercept and slope for the year effect. The model output shows the interaction term as significant. We used the model to predict species richness across years for two latitudinal bands (high and low), and visualised the predictions on the two scales: (*i*) the scale of the response variable, as counts of species (additive), and (*ii*) the scale of the linear predictor used to fit the statistical model (logarithmic, i.e. multiplicative) (Fig. 5).

On both scales, there is a positive trend in site-level species richness over time. On the additive scale, the increase is strongest at low latitudes (as seen by the steeper slope of the purple line in Fig. 5A), indicating that the positive change over time declines with increasing latitude. On the logarithmic scale, by contrast, the increase is stronger for sites at higher latitudes (the yellow line is steeper in Fig. 5B), indicating that the rate of species richness change over time increases with latitude. The direction of the effect modification has reversed across the two scales of measurement. On the logarithmic scale, species richness changes are approximately represented as proportionate changes over time; the low species richness at the beginning of the time series at high latitudes leads to larger proportionate differences over time. Both scales of measurement and analysis can provide important information. Conclusions about variation in the numbers of species redistributions across latitudes would require comparison on the additive scale. In other words, absolute changes in the number of species over time could indicate species shifting their range limits: more species in absolute numbers have shifted their range at lower latitudes. On the other hand, changes on the multiplicative scale could indicate a multiplicative process, for example, the gain of keystone species, which have disproportionate effects on the persistence of other species in an ecosystem.

IV. ASYMMETRIC EXPLORATIONS OF CONTEXT DEPENDENCE ARE INSUFFICIENT TESTS OF THEORIES POSITING INTERACTIONS (TYPE-A ERROR)

Ecological studies often focus on asymmetric hypotheses about context dependence, distinguishing between a focal effect of interest X and a modifier variable Z, and testing for detectable modification of the effect of X by Z. For example, we might ask: how do temporal biodiversity trends vary with biome? Are relationships between biodiversity and ecosystem functioning modified by environmental drivers? Do farming impacts on biodiversity depend on landscape structure? This dichotomy is often justifiable on practical grounds, e.g. because X is a variable that we humans can manipulate (e.g. a conservation intervention), or it represents a given change at a locality, or because a study's sampling strategy (e.g. blocking or randomisation) has been designed for a 'treatment' variable X (Cox, 1984), while Z is a contextual variable that we cannot change, or is outside of the investigator's control (Cox, 1984), such as biome, latitude, taxon, age, rainfall, etc., or an intrinsic variable such as sex. As a result, ecologists tend to visualise, interpret and interrogate statistical interactions asymmetrically.

(1) Visualising interaction effects using marginal effect plots

A common practice to visualise interaction effects is to construct a marginal effect plot that displays how the marginal effect of X on Y (the response coefficient) changes over the range of moderator Z, with all other covariates held constant. Although it is possible to produce two marginal effect plots for an interaction, with Z as moderator of X and *vice versa*, this is rarely done (Berry *et al.*, 2012). The focus on one marginal effect plot, examining effect modification in a single direction can mislead interpretation because any observed relationship between Z and the marginal effect of X could be consistent with multiple underlying conditional relationships. That is, any observed relationship between Z and the marginal effect of X is always consistent with multiple ways in which the marginal effect of Z varies with X, some of which may be inconsistent with the underlying conditional theory being tested (Berry *et al.*, 2012).

(*a*) *Hypothetical example: biodiversity moderates the influence of environmental stress on ecosystem functioning*

We demonstrate the Type-A error using a hypothetical example of the relationship between biodiversity and

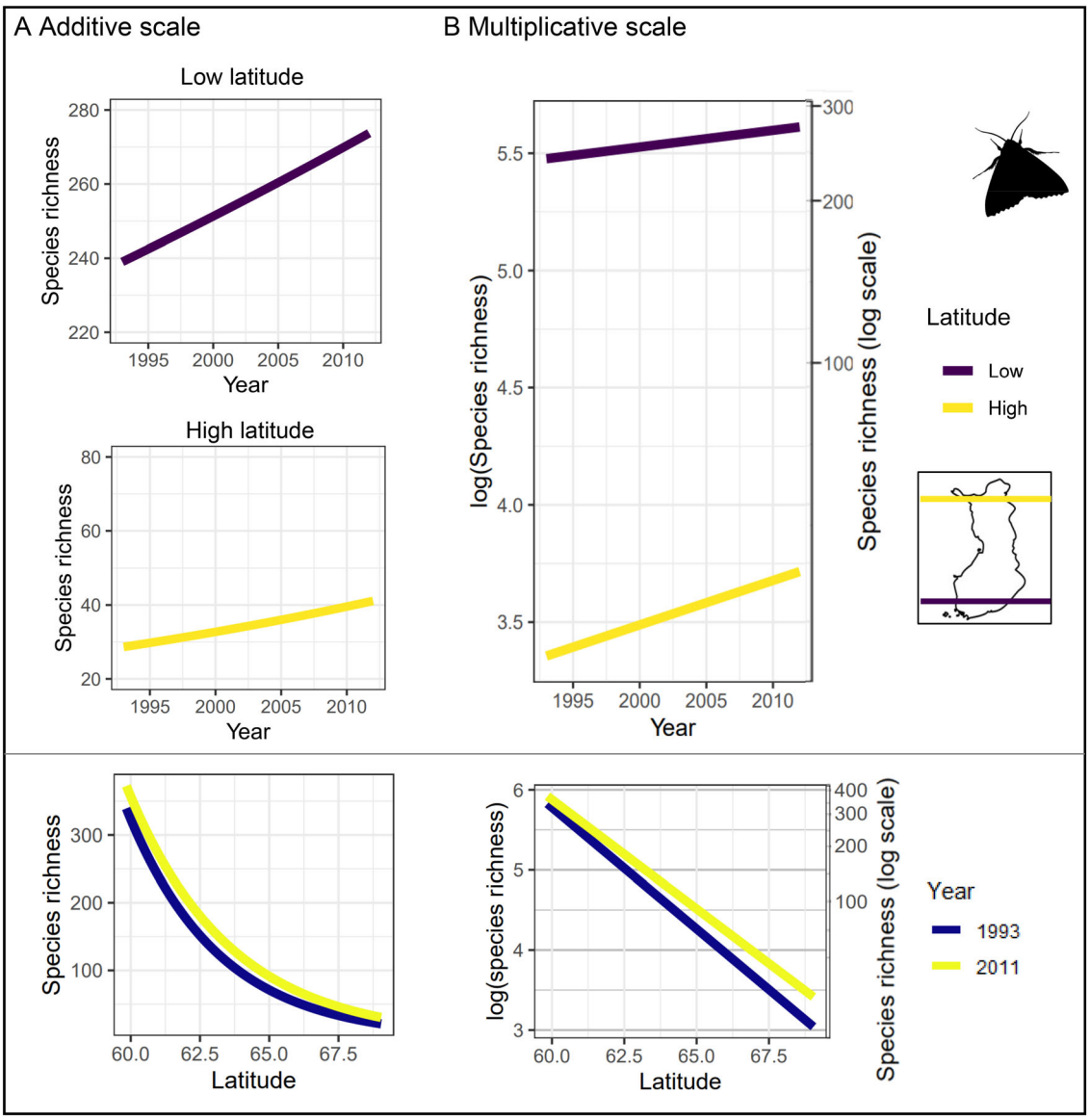

Fig. 5. Top: predicted changes in site-level species richness of moths over time at low (61°, purple) and high (68°, yellow) latitudes in Finland, plotted on (A) additive and (B) multiplicative scales. On the additive scale, the rate of species-richness change over time – the slope of the lines – is stronger at low latitudes (purple line), where sites experienced higher gains in the absolute numbers of species during the monitoring period. By contrast, the rate of richness increases is greatest at high latitudes when presented on the logarithmic scale (yellow line), corresponding to the scale of measurement of a generalised linear mixed model applied to species counts. Bottom: Predicted changes in site-level species richness across latitude shown for 2 years. Data from Antão *et al.* (2020).

ecosystem functioning. Decades of research have demonstrated that biodiversity promotes the functioning of ecosystems (e.g. Hooper *et al.*, 2005; Tilman, Isbell & Cowles, 2014). Studies have sought to identify whether biodiversity can moderate the effect of environmental stress on ecosystem functioning (Tilman *et al.*, 2001), and whether richer communities are more resistant to stress (e.g. Steudel *et al.*, 2012; Baert *et al.*, 2018; Benkwitt, Wilson & Graham, 2020; Hong *et al.*, 2022). In such studies, rather than generating predictions according to a hypothesised causal model, it is common to develop hypotheses that designate biodiversity or ecosystem functioning as a focal variable, and the other as a 'moderator'. Then, the typical approach is to examine asymmetrically how the slopes of the focal driver vary with the moderator variable, to ask whether effects are weakened or strengthened in magnitude across its range.

Consider the following hypothesis of a weakening effect: environmental stress reduces ecosystem functioning, but biodiversity can buffer against this impact. In other words, we expect to see weaker effects of stress on functioning in richer

communities. To test this hypothesis, we identify environmental stress as the focal variable X, and biodiversity as the moderator Z, which weakens the effect of stress on ecosystem functioning Y. We fit a linear model to data (e.g. from a distributed experiment), and specify an interaction term (stress × biodiversity) to represent this hypothesis. After detecting a statistical interaction, we construct a marginal effect plot displaying the estimated effect of stress (the slope), and its change with biodiversity (Fig. 6C).

Fig. 6 shows how an apparent marginal effect trend is consistent with two different scenarios corresponding to different underlying processes. In both scenarios, consistent with our general hypothesis, we observe a weakening trend in the marginal effect of stress with biodiversity – the effect of stress on ecosystem functioning is more weakly negative at higher biodiversity (Fig. 6C). We conclude that highly diverse communities are more resistant to environmental change, and promote management interventions that enhance biodiversity in all contexts (e.g. planting mixtures rather than monocultures). However, a much richer interpretation is gained when we look at the interaction symmetrically, and produce a second marginal effect plot displaying the conditional effects of biodiversity with increasing stress (Fig. 6D), as well as conditional plots displaying predicted levels of ecosystem functioning at high and low stress and biodiversity levels (Fig. 6A, B). Doing so reveals that at low stress, while functioning is relatively high overall, functioning declines with biodiversity in one of the scenarios (bottom of Fig. 6A). In this scenario, functioning would be higher in monocultures in low-stress environments.

The key point here is that marginal effect plots for X (Fig. 6C) do not convey any information about the magnitude or sign of the marginal effect of Z at any value of X. This is critical, because different values for this intercept (in the marginal effect plot) imply very different ways in which the marginal effect of Z is conditional on X, and only some of these ways may be consistent with our theories and hypotheses (Berry *et al.*, 2012). The same patterns can arise from alternative pathways by which X and Z together affect Y. Hence, if we only examine the marginal effect in one direction, we build an incomplete picture of the underlying processes and risk seriously misunderstanding the management implications of the evidence.

V. META-ANALYSIS IS ESPECIALLY VULNERABLE TO ALL THREE INFERENTIAL ERRORS

The systematic collation of studies addressing similar questions, and the subsequent analysis of their summary statistics using meta-analysis, is an increasingly popular approach to

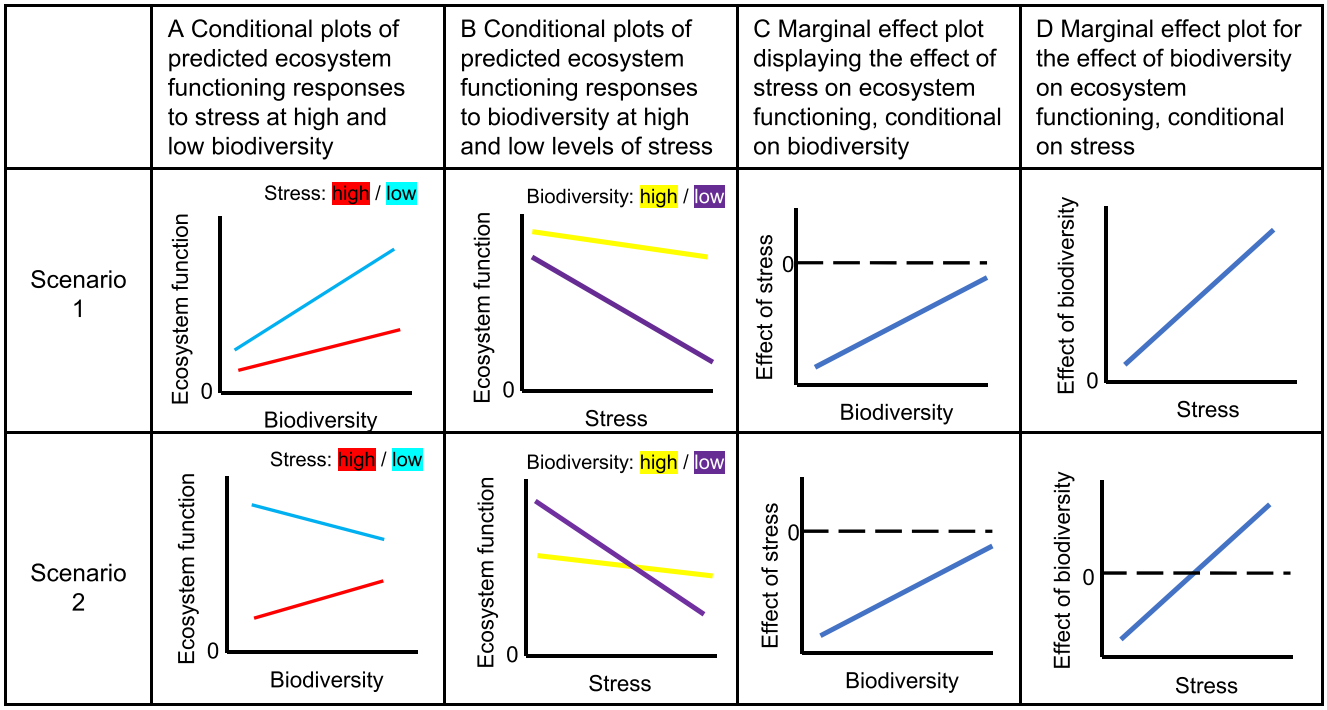

Fig. 6. Two alternative scenarios (rows) of conditional relationships among ecosystem functioning, environmental stress and biodiversity. Marginal effect plots in (C) depict the same weakening effect of stress on ecosystem functioning in more diverse communities. A richer interpretation is gained when we examine the interaction symmetrically, by producing a second marginal effect plot displaying the conditional effect of biodiversity with stress (D), and conditional plots (A, B) that display predicted values of ecosystem functioning at specified values of biodiversity and stress. If we do not examine the relationship symmetrically, we cannot test alternative theories about the interaction.

seeking general patterns in ecology (Anderson *et al.*, 2021). Whilst meta-analysis is often used to ask questions about mean effects, it is also widely promoted as a means of understanding the context dependence of ecological effects amongst studies (called 'heterogeneity' in meta-analysis; e.g. Gurevitch *et al.*, 2018). The classical approach to meta-analysis generally involves three steps (Spake *et al.*, 2022*b*): (*i*) estimation of study-level and overall mean effect sizes; (*ii*) estimation of heterogeneity statistics (such as I^2 or Q-statistics) that quantify variability in study-level effects; and (*iii*) attribution of effect size heterogeneity to predictors (called 'effect modifiers' or 'moderator variables'; Mengersen *et al.*, 2013). The effect size is predominantly estimated with respect to one focal explanatory variable (e.g. the effect of land use on biodiversity, or biodiversity change over time), rendering the meta-analysis asymmetric, and only permitting the assessment of effect modification in a single direction. For instance, if a meta-analysis explores how the effect of X on Y gets modified by Z, the first step is to calculate the effect size (representing the effect of X on Y), which immediately loses sight of the actual values of X and Y. This loss of information during effect size estimation means that it is not possible to examine how the effect of Z on Y gets modified by X. This makes meta-analysis particularly vulnerable to Type-A errors. Moreover, this loss of information removes effect sizes from their baseline values – the mean values, or intercepts, of individual study reference groups – rendering meta-analysis also vulnerable to D and S inferential errors when baselines vary across studies, and the measurement scale is overlooked.

Effect size metrics vary in their measurement scale: implications for Type-D and -S errors

Effect size metrics measure the magnitude and direction of change in Y either as differences between categorical group means, or as the strength of relationships for a focal driver measured on a continuous scale. Effect sizes are considered useful because they allow the collation of data from primary studies that may use different units of measurement (Rohrer & Arslan, 2021). For example, abundance might be measured in counts of individuals, or in biomass units across studies. There are different possible effect size families to choose from; for instance, the *d* family (metrics such as standardised mean difference or Hedges' *g*), or the ratio family (metrics such as the odds ratio or log ratio). From these families, the two most commonly used effect size metrics in ecological meta-analyses are Hedges' $g = [Y_{X2} - Y_{X1}]/SD_{\text{pooled}}$, and the log ratio $= \log[Y_{X2}] - \log[Y_{X2}]$, where Y_{X1} and Y_{X2} are mean outcome values for two levels of X, and SD_{pooled} is the pooled standard deviation of the two groups. There have been several demonstrations of how these metrics vary in their susceptibility to bias under different sampling parameters such as sample size and aerial extent (e.g. Lajeunesse, 2015; Hamman *et al.*, 2018; Spake *et al.*, 2021*a*). Here we examine how these metrics vary in their measurement scale, and discuss the implications for Type-D and -S errors.

These alternative effect size metrics use inherently different scales of measurement, and can thus lead to D and S errors if this is overlooked. In published meta-analyses, the choice of metric is typically justified by the nature and availability of data, rather than the meaningfulness of ecological interpretation (Spake & Doncaster, 2017). For example, Hedges' *g* might be chosen because the presence of zero values renders multiplicative scales uninterpretable and precludes the estimation of log ratios, while the log ratio might be chosen because of unreported SD values that are required for calculating Hedges' *g*. However, the choice of metric has important implications for interpretation. For a normally distributed variable, Hedges' *g* quantifies change on the additive scale (in units of SDs), while the log ratio quantifies multiplicative change and approximates percentage change when effects are small. Regardless of which metric is used, the analyst usually interprets the existence and sign of effect modification without reference to the measurement scale, making such inferences vulnerable to the D and S errors discussed above.

(*a*) *Simulated example: temporal biodiversity trends in actively and passively restored plots following a disturbance event (Type-D and -S error)*

We simulated three data sets to demonstrate the dependence of meta-analytic inference on the choice of effect size metric using R (v4.1.1, R Core Team, 2021), package *AHMbook* (Kéry, Royle & Meredith, 2021); see online Supporting Information, Appendix S1, for details. Each data set represented an independent meta-analysis for a particular taxonomic group, comprising data that had been collated from numerous individual 'studies'. For each data set, we assumed a scenario where species abundances were tracked in multiple plots following a major disturbance event, and each study represented a different point in time since restoration. Replicates of plots were either subjected to active restoration treatment, or left as unrestored control plots (Fig. 7, column A). The taxonomic groups differed in their responses to restoration. For each taxon, we used mean abundance values in restored and control plots to calculate effect sizes that represented the effect of active restoration on abundance for each study, with the metrics: mean difference (the absolute difference between group means), Hedges' *g*, and log ratio (Fig. 7, columns B–D).

All three taxa increased in numbers of individuals through time in both restored and control plots (Fig. 7A), and the rate of increase was faster for actively restored compared to control plots (positive trends in mean difference in Fig. 7B). However, the magnitude and sign of the difference between control and restored plots depends on the effect size metric. For taxa 2 and 3, log ratios show the opposite trend to mean differences, with the positive effect declining with time since disturbance (Fig. 7D). The negative log ratio trend with time might lead the analyst to infer that passively restored sites catch up with actively restored sites given enough time, despite the mean difference increasing over time. For taxon 3, Hedges' *g* remains relatively constant with time since disturbance (Fig. 7C, bottom), because the increasing variability

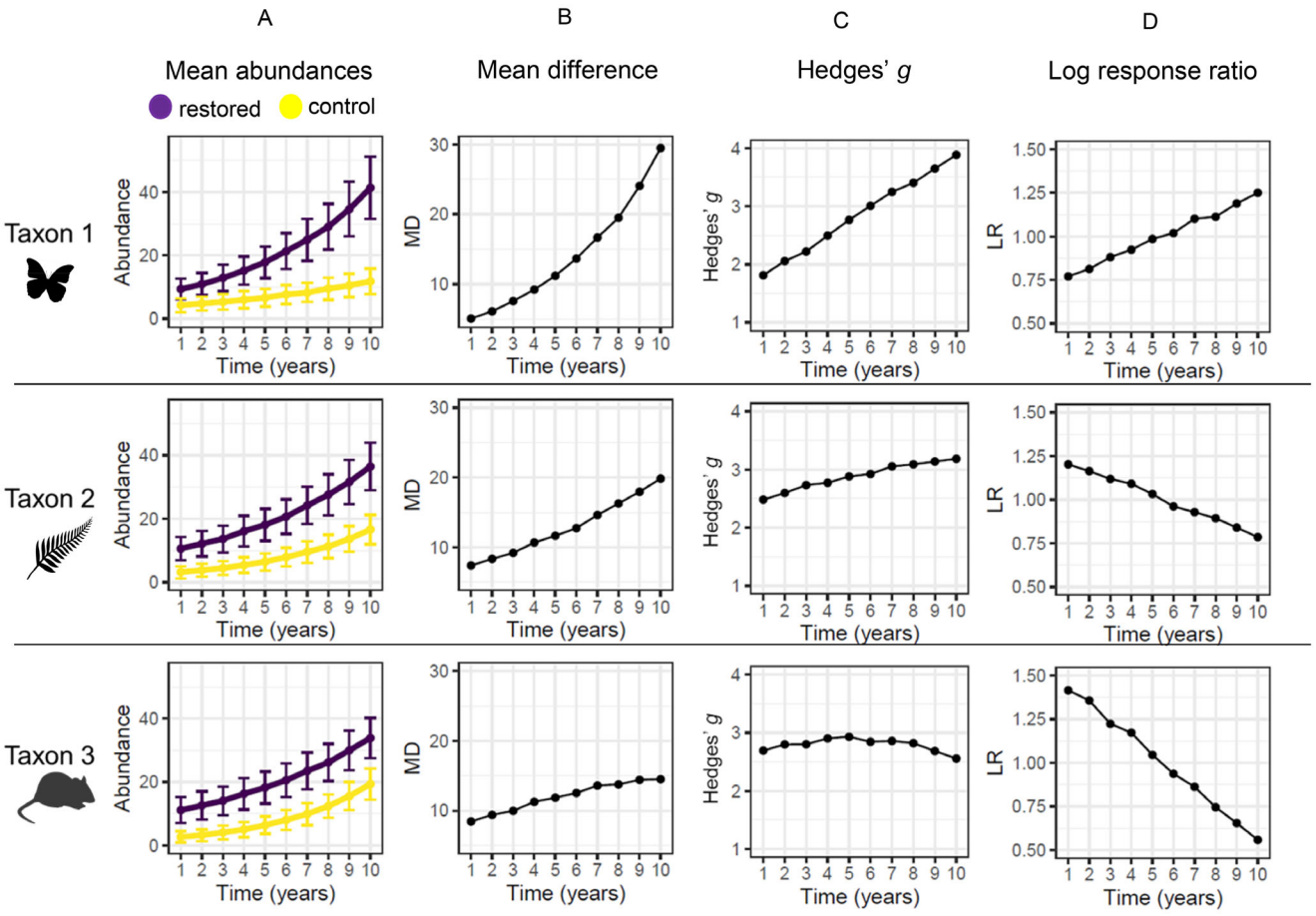

Fig. 7. The magnitude and sign of effect size trends can depend on the effect size metric. We estimated effect sizes and their standard errors from three simulated meta-analytic data sets (see Section V.1.*a*), corresponding to three different taxa (rows). Study-level differences in abundance are shown between actively restored (purple) and control sites (yellow), across time since a major disturbance event (column A). For each meta-analytic data set, we calculated three effect size metrics to represent study-wise differences between actively restored and control sites: mean difference (MD, column B), Hedges' *g* (C), and log ratios (LR; D). For all taxa, the analyst might conclude that the 'effect of restoration gets larger with time since disturbance' for mean difference (B). The positive mean–variance relationship (shown by increasing error bars with mean abundance in column A) can weaken the trend for Hedges' *g* compared to mean difference (e.g. taxon 3 shows a positive effect in B, but C shows no trend). The trend can also reverse in sign with effect size metric, with log ratios measuring proportionate differences (as for taxon 2).

in abundance associated with the increasing mean abundance (as shown by error bars in Fig. 7A), balances out the weaker effect of the increasing abundance difference. This clearly demonstrates the issue that Hedges' *g* is not suited to expressing differences between variables that trend in their mean–variance relationships (Sun & Cheung, 2020).

(*b*) *Empirical example: understorey plant richness differences between managed and unmanaged forests across two continents (Type-D and -S errors)*

Here we demonstrate the influence of effect size metric on inference, with a meta-analysis of data collated by Chaudhary *et al.* (2016) on studies that measured the impacts of forest management on species richness across management types and biomes. We used mean species richness values from unlogged and logged forest plots to calculate four metrics of effect sizes that represent the effect of forest logging on understorey plant species richness: mean difference, Hedges' *g*, log ratio and percentage difference. For each effect size metric, we calculated effect sizes for each primary study, and pooled effect sizes for Europe and North America. This reflects common practice in ecological meta-analysis to estimate overall mean effect sizes across heterogeneous groupings of studies (Senior *et al.*, 2016).

We find that the magnitude of logging effects on understorey richness, and the relative ranking of mean effects by continent (i.e. the sign/direction of effect modification; Fig. 8, right column), vary with the effect size metric (Fig. 8 rows). For the mean difference, the effect of logging is more strongly negative in North America than Europe (Fig. 8C); this was driven by large effect sizes in studies with higher

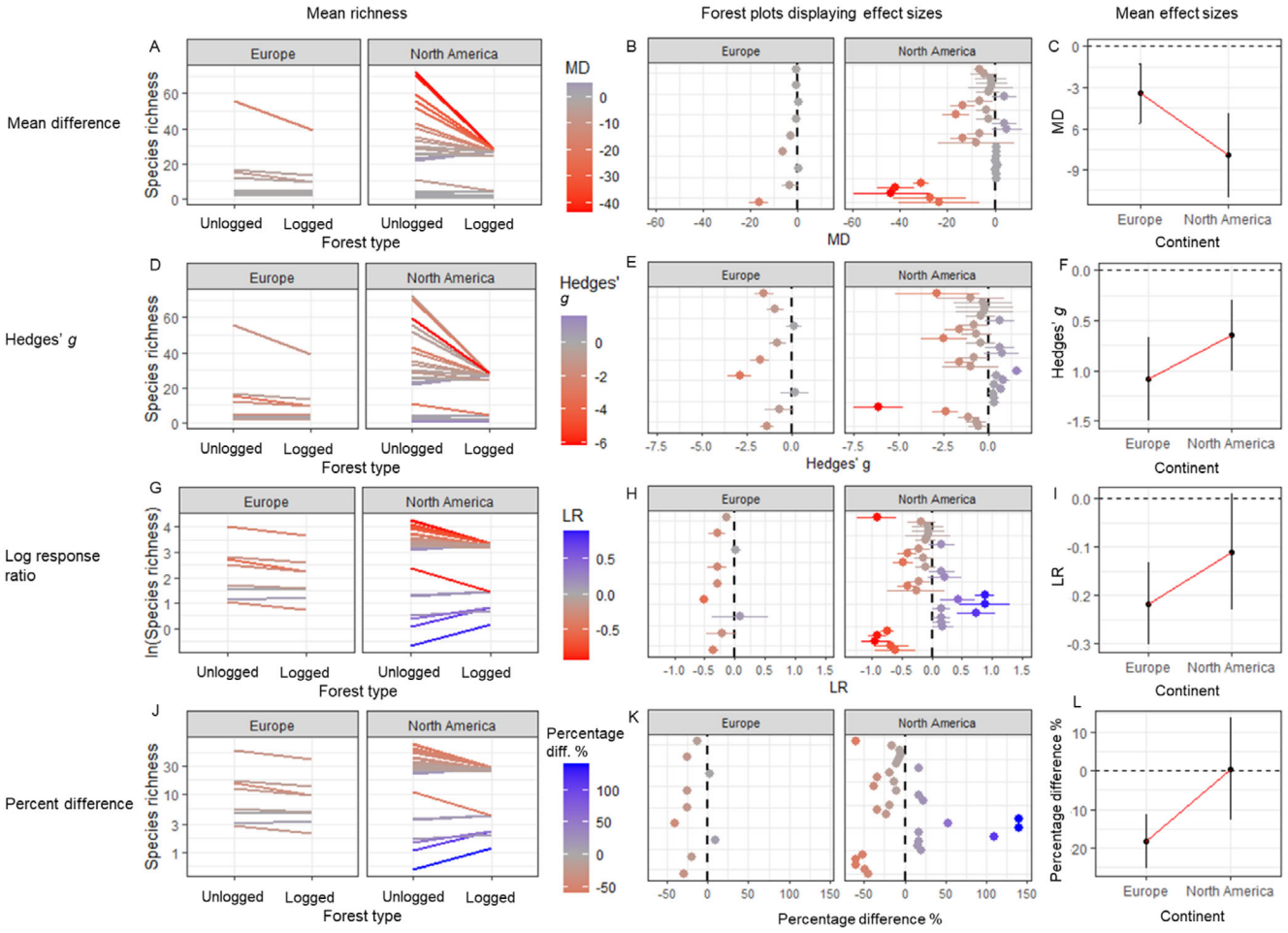

Fig. 8. Interpretation of trends in forest-management impacts depends on effect size metric. Each row corresponds to a different effect size metric and scale of measurement: mean difference (MD), Hedges' *g*, log ratio (LR) and percentage difference. In the left column (A, D, G, J), mean plant richness is shown for studies that compared logged and unlogged forest types across two continents (the same data are plotted but on different scales or coloured by different effect size metrics; each line corresponds to a single study). The middle column of forest plots (B, E, H, K) display effect sizes for individual studies, ordered by study ID and coloured by their values. The right column shows the meta-analysis estimated mean effects of logging by continent, estimated using an unweighted meta-analysis (C, F, I, L). Data compiled by Chaudhary *et al.* (2016).

mean richness that were more common in North America. By contrast, the effect is more strongly negative for Europe using all the other metrics (Fig. 8F, I, L). The difference is not significantly different from zero for North America for the log ratio (I) or percentage difference (L), due to some strongly positive effect sizes on these relative scales from studies with low mean richness (dark blue in Fig. 8G, H, J, K) that balance out negative effects. Our inference therefore depends on the choice of effect size metric.

Why do these differences arise in the magnitude and sign of effect modification? The 'baseline' biodiversity values of the unlogged forest stands (controls) vary widely. The difference in effect size trends between mean differences and percentage change occurs for purely numerical reasons: absolute differences will diverge from ratio differences when baselines vary. The difference in trend between log ratios and percentage change arises because log ratios approximate percentage change only when percentage change is relatively small (as shown by near-zero effects following the 1:1 lines in Fig. 9). Therefore, the log ratio cannot meaningfully represent proportional differences when percentage differences are large, where estimates may exert undue influence on mean effect sizes that are estimated across highly heterogeneous study pools. Large proportionate changes are observed when group mean values are near zero, where any absolute increase in Y becomes large in proportionate terms, and small differences in baseline level lead to drastically different effect size magnitudes (Pustejovsky, 2018). For example, it makes little sense to equate a change from two individuals to four individuals with a change from 102 individuals to either 104 (i.e. +2) or 204 (i.e. ×2) individuals.

Epidemiologists also face the challenge of varying baselines for inferring effect modification from meta-analyses (Chaimani, 2015; Shrier *et al.*, 2016; Yates & Cochran, 1938).

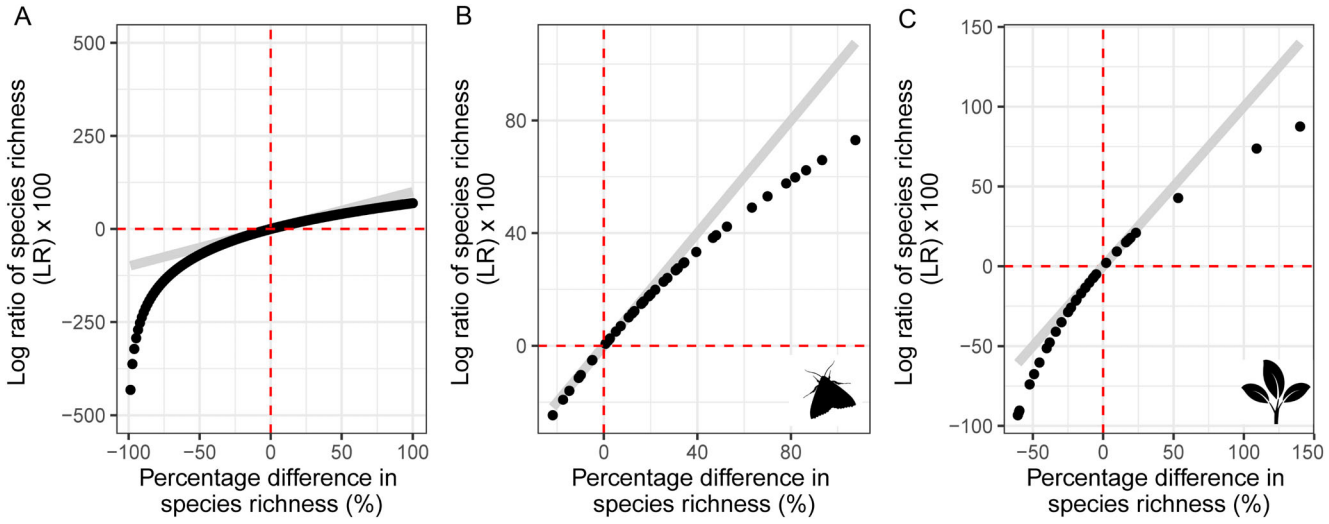

Fig. 9. Correspondence between percentage differences in species richness (*x* axes) and log ratios (LR) multiplied by 100. Effect sizes representing species richness differences are shown for (A) simulated communities with 'control' richness values of 50 and 'treatment' values ranging from 1 to 100; (B) moth communities at the beginning and end of a time series (data from Antão *et al.*, 2020); and (C) understorey plant communities in logged and unlogged forests (data compiled by Chaudhary *et al.*, 2016). Grey lines correspond to a 1:1 match between percentage differences and 100 × LR. Correspondence is greatest when absolute percentage difference is relatively small, at less than ∼±50%. Large positive percentage changes are relatively less strongly expressed as log ratios, while large negative percentage changes are relatively more strongly expressed as log ratios.

For example, in meta-analyses of drug effects on disease risk, differences in 'underlying risk' are important in determining the degree of effect modification by risk factors, as inferred from meta-regression or subgroup analysis. For example, if synthesising studies to compare the effect of an anti-cancer drug on morbidity across different subgroups that vary in average age, the 'baseline' outcome (i.e. morbidity) here covaries with the effect modifier of interest (age). Proposed solutions include using underlying risk as an effect modifier, or measuring change in meaningful, additive units from the baseline (Shrier *et al.*, 2016).

It is worth noting that some meta-analyses comparing multiple effect size metrics have reported similar relationships of effect size moderators, even though these metrics differ as to whether they are additive (Hedges' *g*) or multiplicative (the log ratio). For example, Powell *et al.* (2011) studied the effects of invasive plants on species richness, finding that Hedges' *g* and the log ratio gave similar trends in effect size modification by study spatial extent. We might expect to observe similar trends when the response variable of interest is Poisson-distributed, with a variance that increases with the mean. Hedges' *g* uses the pooled standard deviation to standardise the metric, which increases with the mean, and can cause the metric to have similar behaviour to the log ratio. This similarity only demonstrates that Hedges' *g* is not fit for its purpose of representing additive change for a variable with a significant mean–variance relationship.

(c) *A note on transformation bias*

To improve interpretability, mean log ratio (LR) values are often transformed to percentage change: 100 × [exp (LR) − 1], as a familiar and readily interpretable conceptualisation, which is consistent with how biodiversity scientists and policymakers might discuss biodiversity change. This repurposing of the effect size risks transformation-induced bias, which occurs because a non-linear transformation of a mean value is generally not equal to the mean of transformed values. This is an expression of Jensen's inequality: $f[\mathrm{E}(Y)] \neq \mathrm{E}[f(Y)]$ for an arbitrary random variable Y and non-linear function f (e.g. Nakagawa, Johnson & Schielzeth, 2017). Accordingly, back-transforming the mean value of a log ratio calculated across study-level log ratios introduces a bias into the estimate of the mean percentage difference, due to the convexity of the log transformation. The magnitude of the bias increases with the variance of the weighted mean, which is small only when the number of studies and their precision is high (Hedges, Gurevitch & Curtis, 1999). In ecology, this variance is typically large (Senior *et al.*, 2016), and can vary widely across subgroupings of studies. A potential solution to this problem for approximately normally distributed data is to use a correction factor: $100 \times [\exp(\mathrm{LR} + 0.5 \times V_{\mathrm{total}}) - 1]$, where V_{total} is the variance of all log ratio values (Nakagawa *et al.*, 2017).

VI. GUIDELINES TO IMPROVE INFERENCE ABOUT CONTEXT DEPENDENCE

Given that quantifying context dependence will remain a major focus across theoretical and applied ecology despite the potential for D, S and A errors described above, we provide guidance below in the form of numbered points to

improve inference, focusing on hypothesis generation, modelling considerations, and visualising and interpreting context dependence.

(1) Hypothesis generation

(1) Hypotheses and predicted patterns should be aligned clearly to causal models. Epidemiologists often distinguish between 'effect measure modification' (Rothman, 2002), where magnitude or sign of the effect of X on Y (on a particular measurement scale) varies with the level of a third variable Z, where the effect of Z may or may not be causal, and 'biological interaction', denoting the interdependent, reciprocal, or mutual operation, actions, or effects of X and Z on Y, where relationships with X and Z are both causal (Vanderweele, 2009; Bours, 2021). We do not wish to dictate terminology, but instead emphasise the importance of *a priori* causal reasoning.

(2) Hypotheses and predicted patterns should be aligned clearly to additive and/or multiplicative processes, where possible. If the scale of relevance is unclear, hypotheses could be made on both scales. Table 2 includes examples of scales of interest for both theoretical and applied questions, and whether they correspond to the scale of modelling (see point 11). The most important consideration is to distinguish effect modification that arises only from ceiling and floor effects of biological phenomena from effect modification that arises from other biological mechanisms that would still be interactions on additive scales. For example, either cold or starvation can kill an animal. Thus, temperature and resource availability must modify each other's effects on survival on the multiplicative scale, even if they do not on the additive scale. But they might also modify each other on the additive scale if, for example, it is easier to starve when conditions are cold. Essentially, an animal can only die once, forcing a log-linear scale, and statistical interactions therefore do not necessarily imply a biological interaction.

Table 2. Examples of theoretical and applied reasons for selecting additive or multiplicative scales for common ecological response variables.

Response variable	Type of modelling	Scale of interest for hypothesis generation, visualisation and interpretation	
		Theoretical context	Applied context
Species richness (counts)	Poisson or negative binomial with logarithmic link function; or log transformation prior to linear modelling	*Additive*: forecasting net gains and losses over time due to (local) extinctions or to species' redistributions brought about by range shifts, e.g. due to colonisations from an external species pool *Multiplicative*: species interactions or trophic cascades, where species have disproportionate influences on ecosystem structure; for example, diversity begetting diversity	*Additive*: interventions aiming to maximise absolute numbers of species, regardless of baseline species richness; interventions targeted to areas that will generate the largest absolute effect on richness. *Multiplicative*: identifying where effects are proportionately greatest, for similar baselines and constrained range of effects (e.g. 20–60% differences)
Abundance (counts)	Poisson or negative binomial with logarithmic link function; or log transformation prior to linear modelling	*Additive*: predicting net gains and losses over time due to redistributions or changes to activity patterns *Multiplicative*: identifying multiplicative processes such as population growth, and *per capita* effects on activity	*Additive*: contributions that change linearly with abundance (e.g. provisioning ecosystem service; invasive species) *Multiplicative*: drivers of species' long-term population trends
Occurrence (binary)	Binomial family with logit link function	*Additive*: rare in practice *Multiplicative*: predicting species distributions based on environmental covariates, for prediction probabilities bounded between 0 and 1	*Additive*: change in probability, or 'risk difference', for rare species of conservation concern, or invasive species targeted for eradication *Multiplicative*: as above, but change in the relative risk might be informative for species that are not rare, for which probability does not approach zero
Proportion of infected individuals (proportion)	Binomial family with logit link function	*Additive*: thresholds of herd immunity *Multiplicative*: forecasting proportions under different scenarios	*Additive*: thresholds of herd immunity *Multiplicative*: ceiling effects if total eradication is of interest
Proportion of individuals surviving a stage or time interval	Binomial family with logit link function	*Multiplicative*: assessment of multiple-predator effects on prey survival	*Additive*: changes to number of surviving adults of rare species per unit effort in response to multiple stressors such as pollution and temperature

(3) Make symmetric predictions about effect modification not only on the modification of X effects by Z, but also the modification of Z effects by X. Be aware that testing a statistical interaction involves multiple hypotheses that can be unpacked to increase the strength of inferences drawn from the study [see Berry *et al.* (2012) for detailed guidance]. The crucial issue is that any contingent association between two variables can arise from multiple causal mechanisms. These multiple mechanisms matter when extrapolating or trying to transport effects across studies (Spake *et al.*, 2022*b*). Tests of conditional theories should be informed by *a priori* causal theory where possible.
(4) To avoid asymmetry and encourage more nuance when testing theories, analysts could construct hypothetical conditional plots: graphical displays of the predicted values of Y at minimum, maximum and/or substantively meaningful values of both X and Z (Berry *et al.*, 2012).

(2) Statistical modelling

(5) Use an error structure that matches the biological process being modelled (e.g. Kerkhoff & Enquist, 2009; Cawley & Janacek, 2010), or appropriately transform model predictions to the scale of interest if alternative error structures are required as ascertained by statistical analysis (Xiao *et al.*, 2011). The appropriate functional form might be evaluated by exploratory scatterplots or inspections of residuals from preliminary models. The choice of scale might be influenced by the range over which the response values vary. For example, when modelling a proportions data set that are largely in a middle range (0.3–0.7), a linear scale might be better than a logit scale. See Table 2 for examples of scales of measurement and scales of interest for common response variables in theoretical and applied ecology.
(6) Synthetic studies that analyse raw study-level data are preferred to meta-analyses of study-level summary data, when possible. Analyses of raw data can allow a more complete test of interactions, because meta-analyses of effect sizes inherently impose an asymmetry and divorce the analyst from baselines.
(7) For meta-analyses, be aware that the magnitude and sign of an effect size trend depends on the effect size metric used, due to influences of data distribution (non-normality) and/or heterogeneity of variances, and differences in baseline values. Do not use Hedges' g with Poisson-distributed variables due to its standardisation by SD_{pooled}, which can covary with the mean. Log ratios as a proportionate measure of change cannot meaningfully represent effect sizes when comparing groups with near-zero means or with large differences in baseline (e.g. control group) values between studies.
(8) Be aware of potential transformation biases when transforming averaged model predictions and use appropriate corrections.

Visualisation and interpretation

(9) Any statement about context dependence being 'stronger' or 'weaker' in different contexts, must be scale specific (i.e. whether the relative magnitude or existence of context dependency exists on a multiplicative or additive scale). Be aware that statistical interaction indicates departure from the underlying form of a fitted statistical model (Rothman, 2002), such that the effect of each explanatory variable on the response varies with the magnitude or sign of other influential variables. Detection of a statistical interaction therefore does not necessarily imply a biological interaction, for example if the interaction is enforced by ceiling/floor constraints on the response variable.
(10) Graphical displays are essential to the interpretation and communication of context dependence. If uncertain about the relative importance of additive and multiplicative processes, visualise and interpret model predictions on both measurement scales (i.e. scale of model and transformed predictions).
(11) Marginal-effect plots that display predicted coefficients of X as conditional on values of Z are asymmetric and omit information about the observed data underlying an interaction (i.e. baselines). Where possible, analysts should instead or additionally use conditional plots that display predicted values of Y across substantively meaningful values of both X and Z (e.g. using faceting or three-dimensional plots). These graphs can then be compared with the predicted relationships to evaluate whether intercepts and slopes are consistent with hypothesised relationships. The scale at which results are presented and communicated might be different to the scale used for modelling, and this should be made clear when describing the analysis and findings.
(12) Display conditional plots for generalised linear models with non-linear link functions, even without interaction terms, because they are inherently multiplicative and therefore interactive. Graphically assess effect modification from generalised linear models even if the interaction term is non-significant (Rönkkö *et al.*, 2022).
(13) When interpreting published research, be aware of the types of interactions that are particularly vulnerable to inferential errors. Appendix S2 provides examples of statistical interactions and their vulnerabilities to Types D, S and A errors.
(14) Seek to move beyond static two-dimensional graphical displays for communicating context dependence. Many disciplines increasingly use interactive web applications that enable the generation of predictions for user-specified inputs (McCabe, Kim & King, 2018; Perkel, 2018; Weissgerber *et al.*, 2019; in ecology: Spake *et al.*, 2020). Such applications enrich understanding of the scale and symmetry of interactions by allowing users to interact directly with underlying data, and choose which variables and on which measurement scale to plot predictions.

VII. CONCLUSIONS

(1) Ecologists routinely use statistical models to detect and explain interactions amongst ecological drivers, with a goal

to evaluate whether an effect of interest changes in sign or magnitude in different contexts. Three common inferential errors arise when ecologists interpret statistical interactions without paying attention to their fundamental property of symmetry, or to the measurement scale, whether additive or multiplicative. These errors take three principal forms: failing to detect ('D' errors), and mistaking the sign ('S' errors) of the dependency, and misidentifying the underlying causal model ('A' errors).

(2) Meta-analysis, which has become a widely used tool for characterising context dependence in ecology, is especially prone to all three errors. The magnitude and sign of an effect size trend depends on the effect size metric used, due to differences in their scale of measurement (whether additive or multiplicative), influences of data distribution, and differences in baseline values. Future syntheses should prioritise full analysis of raw data over meta-analysis of summary statistics. If only meta-analysis is possible, researchers must justify their choice of effect size metric with respect to ecological interpretation.

(3) Symmetry and the interaction scale must be considered explicitly during hypothesis generation, testing, visualisation and interpretation of context dependence in ecology.

(4) While our review has focused on issues most pertinent to common types of ecological data, our article serves as a starting point for improving present practices in hypothesis generation, modelling and visual display of interactions in ecology.

VIII. ACKNOWLEDGEMENTS

R. S. was funded by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. C. T. C. was supported by a Marie Skłodowska-Curie Individual Fellowship (no. 891052). L. H. A. was funded by the Academy of Finland (grant 340280). We thank I. Oliver for supplying the ant data for Fig. 4. We thank D. Craven for motivating Figure 9 with their blog post on nonlinear properties of response ratios. We are grateful to B. Bolker and N.G. Yoccoz for reviewing and improving an earlier version of this manuscript.

IX. REFERENCES

References identified with an asterisk (*) are cited only within the supporting information.

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights* **2**, 193–208.

Albertson, L. K., MacDonald, M. J., Tumolo, B. B., Briggs, M. A., Maguire, Z., Quinn, S., Sanchez-Ruiz, J. A., Veneros, J. & Burkle, L. A. (2021). Uncovering patterns of freshwater positive interactions using meta-analysis: identifying the roles of common participants, invasive species and environmental context. *Ecology Letters* **24**, 594–607.

Anderson, S. C., Elsen, P. R., Hughes, B. B., Tonietto, R. K., Bletz, M. C., Gill, D. A., Holgerson, M. A., Kuebbing, S. E., McDonough MacKenzie, C., Meek, M. H. & Veríssimo, D. (2021). Trends in ecology and conservation over eight decades. *Frontiers in Ecology and the Environment* **19**, 274–282.

Antao, L. H., Pöyry, J., Leinonen, R. & Roslin, T. (2020). Contrasting latitudinal patterns in diversity and stability in a high-latitude species-rich moth community. *Global Ecology and Biogeography* **29**, 896–907.

Baert, J. M., Eisenhauer, N., Janssen, C. R. & De Laender, F. (2018). Biodiversity effects on ecosystem functioning respond unimodally to environmental stress. *Ecology Letters* **21**, 1191–1199.

Benkwitt, C. E., Wilson, S. K. & Graham, N. A. J. (2020). Biodiversity increases ecosystem functions despite multiple stressors on coral reefs. *Nature Ecology and Evolution* **4**, 919–926.

Berry, W. D., Golder, M. & Milton, D. (2012). Improving tests of theories positing interaction. *Journal of Politics* **74**, 653–671.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24**, 127–135.

Bours, M. J. L. (2021). Tutorial: a nontechnical explanation of the counterfactual definition of effect modification and interaction. *Journal of Clinical Epidemiology* **134**, 113–124.

Bradley, M., Nagelkerken, I., Baker, R. & Sheaves, M. (2020). Context dependence: a conceptual approach for understanding the habitat relationships of coastal marine fauna. *Bioscience* **70**, 986–1004.

Catford, J. A., Wilson, J. R. U., Pysek, P., Hulme, P. E. & Duncan, R. P. (2021). Addressing context dependence in ecology. *Trends in Ecology & Evolution* **37**, 158–170.

Cawley, G. C. & Janacek, G. J. (2010). On allometric equations for predicting body mass of dinosaurs. *Journal of Zoology* **280**, 355–361.

Chaimani, A. (2015). Accounting for baseline differences in meta-analysis. *Evidence-Based Mental Health* **18**, 23–26.

Chaudhary, A., Burivalova, Z., Koh, L. P. & Hellweg, S. (2016). Impact of forest management on species richness: global meta-analysis and economic trade-offs. *Scientific Reports* **6**, 1–10.

Cox, D. R. (1984). Interaction. *International Statistical Review* **52**, 1–24.

Duncan, R. P. & Kefford, B. J. (2021). Interactions in statistical models: three things to know. *Methods in Ecology and Evolution* **12**, 2287–2297.

Greenland, S. (2015). Effect modification and interaction. In *Wiley StatsRef: Statistics Reference Online*, pp. 1–5. University of California, Los Angeles, CA.

Griffen, B. D., Belgrad, B. A., Cannizzo, Z. J., Knotts, E. R. & Hancock, E. R. (2016). Rethinking our approach to multiple stressor studies in marine environments. *Marine Ecology Progress Series* **543**, 273–281.

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

Hallqvist, J., Ahlbom, A. & Reuterwall, C. (1996). How to evaluate interaction between causes: a review of practices in cardiovascular epidemiology. *Journal of Internal Medicine* **239**, 377–382.

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

Hedges, L., Gurevitch, J. & Curtis, P. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150–1156.

Hong, P., Schmid, B., De Laender, F., Eisenhauer, N., Zhang, X., Chen, H., Craven, D., De Boeck, H. J., Hautier, Y., Petchey, O. L., Reich, P. B., Steudel, B., Striebel, M., Thakur, M. P. & Wang, S. (2022). Biodiversity promotes ecosystem functioning despite environmental change. *Ecology Letters* **25**, 555–569.

Hooper, D. U., Chapin, F. S. III, Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., Lawton, J. H., Lodge, D. M., Loreau, M., Naeem, S., Schmid, B., Setälä, H., Symstad, A. J., Vandermeer, J. & Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs* **75**, 3–35.

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

Kerkhoff, A. J. & Enquist, B. J. (2009). Multiplicative by nature: why logarithmic transformation is necessary in allometry. *Journal of Theoretical Biology* **257**, 519–521.

Kéry, M., Royle, A. & Meredith, M. (2021). AHMbook: functions and data for the book 'applied hierarchical modeling in ecology' vols 1 and 2. R package version 0.2.3. https://cran.r-project.org/package=AHMbook.

Knol, M. J., van der tweel, I., Grobbee, D. E., Numans, M. E. & Geerlings, M. I. (2007). Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *International Journal of Epidemiology* **36**, 1111–1118.

Knol, M. J. & VanderWeele, T. J. (2012). Recommendations for presenting analyses of effect modification and interaction. *International Journal of Epidemiology* **41**, 514–520.

Knol, M. J., VanderWeele, T. J., Groenwold, R. H. H., Klungel, O. H., Rovers, M. M. & Grobbee, D. E. (2011). Estimating measures of interaction on an additive scale for preventive exposures. *European Journal of Epidemiology* **26**, 433–438.

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

Leal, L. C. & Peixoto, P. E. C. (2017). Decreasing water availability across the globe improves the effectiveness of protective ant–plant mutualisms: a meta-analysis. *Biological Reviews* **92**, 1785–1794.

Leimu, R., Mutikainen, P., Koricheva, J. & Fischer, M. (2006). How general are positive relationships between plant population size, fitness and genetic variation? *Journal of Ecology* **94**, 942–952.

Leinonen, R., Pöyry, J., Söderman, G. & Tuominen-Roto, L. (2016). *Suomen yöperhosseuranta (Nocturna) 1993–2012 [the Finnish Moth Monitoring Scheme (Nocturna) 1993–2012]*. Suomen ympäristökeskuksen Raportteja, 15/2016, Helsinki.

Leung, B., Hargreaves, A. L., Greenberg, D. A., McGill, B., Dornelas, M. & Freeman, R. (2020). Clustered versus catastrophic global vertebrate declines. *Nature* **588**, 267–271.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition* **6**, 312–319.

Loreau, M., Cardinale, B. J., Isbell, F., Newbold, T., O'Connor, M. I. & de Mazancourt, C. (2022). Do not downplay biodiversity loss. *Nature* **601**, E27–E28.

Manfrin, A., Singer, G., Larsen, S., Weiss, N., van Grunsven, R. H. A., Weiss, N. S., Wohlfahrt, S., Monaghan, M. T. & Hölker, F. (2017). Artificial light at night affects organism flux across ecosystem boundaries and drives community structure in the recipient ecosystem. *Frontiers in Environmental Science* **5**, 61.

Marino, N. d. A. C., Romero, G. Q. & Farjalla, V. F. (2018). Geographical and experimental contexts modulate the effect of warming on top-down control: a meta-analysis. *Ecology Letters* **21**, 455–466.

Mayo, D. G. & Hand, D. (2022). Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese* **200**, 220.

McCabe, C. J., Halvorson, M. A., King, K. M., Cao, X. & Kim, D. S. (2022). Interpreting interaction effects in generalized linear models of nonlinear probabilities and counts. *Multivariate Behavioral Research* **57**, 243–263.

McCabe, C. J., Kim, D. S. & King, K. M. (2018). Improving present practices in the visual display of interactions. *Advances in Methods and Practices in Psychological Science* **1**, 147–165.

Mengersen, K., Schmid, C. H., Jennions, M. D. & Gurevitch, J. (2013). Statistical models and approaches to inference. In *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton University Press, Princeton.

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

Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface* **14**, 20170213.

Oliver, I., Dorrough, J., Doherty, H. & Andrew, N. R. (2016). Additive and synergistic effects of land cover, land use and climate on insect biodiversity. *Landscape Ecology* **31**, 2415–2431.

Perkel, J. M. (2018). The future of scientific figures. *Nature* **554**, 133–134.

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

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology* **68**, 99–112.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohrer, J. M. & Arslan, R. C. (2021). Precise answers to vague questions: issues with interactions. *Advances in Methods and Practices in Psychological Science* **4**, 25152459211007368.

Rönkkö, M., Aalto, E., Tenhunen, H. & Aguirre-Urreta, M. I. (2022). Eight simple guidelines for improved understanding of transformations and nonlinear effects. *Organizational Research Methods* **25**, 48–87.

Rothman, K. J. (2002). *Epidemiology: An Introduction*. Oxford University Press, New York.

Rothman, K. J., Greenland, S. & Walker, A. M. (1980). Concepts of interaction. *American Journal of Epidemiology* **112**, 467–470.

Senior, A. M., Grueber, C. E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E. S. A. & Nakagawa, S. (2016). Heterogeneity in ecological and evolutionary meta- analyses: its magnitude and implications. *Ecology* **97**, 3293–3299.

Seufert, V. & Ramankutty, N. (2017). Many shades of gray—the context-dependent performance of organic agriculture. *Science Advances* **3**, e1602638.

Shrier, I., Christensen, R., Juhl, C. & Beyene, J. (2016). Meta-analysis on continuous outcomes in minimal important difference units: an application with appropriate variance calculations. *Journal of Clinical Epidemiology* **80**, 57–67.

Smith, O. M., Cohen, A. L., Reganold, J. P., Jones, M. S., Orpet, R. J., Taylor, J. M., Thurman, J. H., Cornell, K. A., Olsson, R. L., Ge, Y., Kennedy, C. M. & Crowder, D. W. (2020). Landscape context affects the sustainability of organic farming systems. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 2870–2878.

Smokorowski, K. E. & Randall, R. G. (2017). Cautions on using the before-after-control-impact design in environmental effects monitoring programs. *Facets* **2**, 212–232.

Spake, R., Barajas-barbosa, M. P., Blowes, S. A., Bowler, D. E., Callaghan, C. T., Garbowski, M., Jurburg, S. D., van Klink, R., Korell, L., Ladouceur, E., Rozzi, R., Viana, D. S., Xu, W. & Chase, J. M. (2022*a*). Detecting thresholds of ecological change in the Anthropocene. *Annual Review of Environment and Resources* **47**, 797–821.

Spake, R., Bellamy, C., Gill, R., Watts, K., Wilson, T., Ditchburn, B. & Eigenbrod, F. (2020). Forest damage by deer depends on cross-scale interactions between climate, deer density and landscape structure. *Journal of Applied Ecology* **57**, 1376–1390.

Spake, R., Bellamy, C., Graham, L., Watts, K., Wilson, T., Norton, L., Wood, C., Schmucki, R., Bullock, J. & Eigenbrod, F. (2019). An analytical framework for spatially targeted management of natural capital. *Nature Sustainability* **2**, 90–97.

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

Spake, R., Mori, A. S., Beckmann, M., Martin, P. A., Christie, A. P., Duguid, M. C. & Doncaster, C. P. (2021*a*). Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecology Letters* **24**, 374–390.

Spake, R., O'Dea, R. E., Nakagawa, S., Doncaster, C. P., Ryo, M., Callaghan, C. T. & Bullock, J. M. (2022*b*). Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution* **6**, 1818–1828.

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

Spiegelman, D. & VanderWeele, T. J. (2017). Evaluating public health interventions: 6. Modeling ratios or differences? Let the data tell us. *American Journal of Public Health* **107**, 1087–1091.

Steudel, B., Hector, A., Friedl, T., Löfke, C., Lorenz, M., Wesche, M. & Kessler, M. (2012). Biodiversity effects on ecosystem functioning change along environmental stress gradients. *Ecology Letters* **15**, 1397–1405.

Sun, R. W. & Cheung, S. F. (2020). The influence of nonnormality from primary studies on the standardized mean difference in meta-analysis. *Behavior Research Methods* **52**, 1552–1567.

Tilman, D., Isbell, F. & Cowles, J. M. (2014). Biodiversity and ecosystem functioning. *Annual Review of Ecology, Evolution, and Systematics* **45**, 471–493.

Tilman, D., Reich, P. B., Knops, J., Wedin, D., Mielke, T. & Lehman, C. (2001). Diversity and productivity in a long-term grassland experiment. *Science* **294**, 843–845.

Vanderweele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology* **20**, 863–871.

Vanderweele, T. J. (2019). The interaction continuum. *Epidemiology* **30**, 648–658.

VanderWeele, T. J. & Knol, M. J. (2014). A tutorial on interaction. *Epidemiologic Methods* **3**, 33–72.

VanderWeele, T. J. & Robins, J. M. (2007). The identification of synergism in the sufficient-component-cause framework. *Epidemiology* **18**, 329–339.

*Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36**, 1–48.

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H. & Iverson, G. (2012). On the interpretation of removable interactions: a survey of the field 33 years after Loftus. *Memory and Cognition* **40**, 145–160.

Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *American Statistician* **73**, 1–19.

Wauchope, H. S., Amano, T., Geldmann, J., Johnston, A., Simmons, B. I., Sutherland, W. J. & Jones, J. P. G. (2021). Evaluating impact using time-series data. *Trends in Ecology and Evolution* **36**, 196–205.

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D. & Milic, N. M. (2019). Reveal, don't conceal: transforming data visualization to improve transparency. *Circulation* **140**, 1506–1518.

Wirsing, A. J., Heithaus, M. R., Brown, J. S., Kotler, B. P. & Schmitz, O. J. (2021). The context dependence of non-consumptive predator effects. *Ecology Letters* **24**, 113–129.

Wolkovich, E. M., Auerbach, J., Chamberlain, C. J., Buonaiuto, D. M., Ettinger, A. K., Morales-Castilla, I. & Gelman, A. (2021). A simple explanation for declining temperature sensitivity with warming. *Global Change Biology* **27**, 4947–4949.

Xiao, X., White, E. P., Hooten, M. B. & Durham, S. L. (2011). On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology* **92**, 1887–1894.

Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I. & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global Change Biology* **28**, 969–989.

Yates, F. & Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science* **28**, 556–580.

X. SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1. Details of meta-analysis of simulated primary studies measuring biodiversity in actively restored and control forest plots following major disturbance events.
Appendix S2. Examples of statistical interactions and their vulnerabilities to Types D, S and A errors.

(*Received 21 June 2022; revised 4 February 2023; accepted 7 February 2023*)