Check for updates

Received: 3 May 2019 | Accepted: 12 August 2019

DOI: 10.1111/1365-2664.13499

RESEARCH ARTICLE

Journal of Applied Ecology BRITISH ECOLOGICAL SOCIETY

Simple study designs in ecology produce inaccurate estimates of biodiversity responses

Alec P. Christie[1] | Tatsuya Amano[1,2,3] | Philip A. Martin[1,4] | Gorm E. Shackelford[1,4] | Benno I. Simmons[1,5] | William J. Sutherland[1,4]

[1]Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK

[2]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK

[3]School of Biological Sciences, University of Queensland, Brisbane, QLD, Australia

[4]BioRISC, St Catharine's College, Cambridge, UK

[5]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

Correspondence
Alec P. Christie
Email: apc58@cam.ac.uk

Funding information
Natural Environment Research Council, Grant/Award Number: NE/L002507/1; Kenneth Miller Trust; The David and Claudia Harding Foundation; Grantham Foundation for the Protection of the Environment; Australian Research Council, Grant/Award Number: FT180100354; Arcadia

Handling Editor: Júlio Louzada

Abstract

1. Monitoring the impacts of anthropogenic threats and interventions to mitigate these threats is key to understanding how to best conserve biodiversity. Ecologists use many different study designs to monitor such impacts. Simpler designs lacking controls (e.g. Before–After (BA) and After) or pre-impact data (e.g. Control–Impact (CI)) are considered to be less robust than more complex designs (e.g. Before–After Control-Impact (BACI) or Randomized Controlled Trials (RCTs)). However, we lack quantitative estimates of how much less accurate simpler study designs are in ecology. Understanding this could help prioritize research and weight studies by their design's accuracy in meta-analysis and evidence assessment.
2. We compared how accurately five study designs estimated the true effect of a simulated environmental impact that caused a step-change response in a population's density. We derived empirical estimates of several simulation parameters from 47 ecological datasets to ensure our simulations were realistic. We measured design performance by determining the percentage of simulations where: (a) the true effect fell within the 95% Confidence Intervals of effect size estimates, and (b) each design correctly estimated the true effect's direction and magnitude. We also considered how sample size affected their performance.
3. We demonstrated that BACI designs performed: 1.3–1.8 times better than RCTs; 2.9–4.2 times versus BA; 3.2–4.6 times versus CI; and 7.1–10.1 times versus After designs (depending on sample size), when correctly estimating true effect's direction and magnitude to within ±30%. Although BACI designs suffered from low power at small sample sizes, they outperformed other designs for almost all performance measures. Increasing sample size improved BACI design accuracy, but only increased the precision of simpler designs around biased estimates.
4. *Synthesis and applications*. We suggest that more investment in more robust designs is needed in ecology since inferences from simpler designs, even with large sample sizes may be misleading. Facilitating this requires longer-term funding and stronger research–practice partnerships. We also propose 'accuracy weights' and demonstrate how they can weight studies in three recent meta-analyses by

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. *Journal of Applied Ecology* published by John Wiley & Sons Ltd on behalf of British Ecological Society

accounting for study design and sample size. We hope these help decision-makers and meta-analysts better account for study design when assessing evidence.

KEYWORDS
Before–After Control–Impact, causal inference, evidence synthesis, impact evaluation, inverse-variance weighting, meta-analysis, randomized controlled trial, study design

1 | INTRODUCTION

Monitoring the impact of human activities on biodiversity is fundamental to understanding how to effectively conserve biodiversity. This includes monitoring the impacts of anthropogenic threats, as well as the effectiveness of management interventions to mitigate such threats. The main challenge for such monitoring is disentangling natural environmental change from anthropogenic change (Hewitt, Thrush, & Cummings, 2001; Hipel, Lettenmaier, & McLeod, 1978), while considering the focal impact's statistical (Box & Tiao, 1975; Osenberg & Schmitt, 1996) and ecological significance (Wolfe, Champ, Flemer, & Mearns, 1987). The complexity of ecosystems, including various sources of spatiotemporal variation and confounding variables, has catalysed much research on understanding the best ways to design impact assessments (Lettenmaier, Hipel, & McLeod, 1978; Osenberg, Bolker, White, St Mary, & Shima, 2006; Stewart-Oaten, Murdoch, & Parker, 1986). While improvements in study design have helped ecologists to more accurately quantify human impacts on biodiversity, a range of designs with varying complexity and biases still persist (De Palma et al., 2018; Figure 1).

Study design is composed of three major aspects: (a) pre-impact sampling, (b) use of controls, and (c) randomized allocation of independent sampling units (here we term these 'sites'). Adding pre-impact sampling to an After design – where monitoring only occurs after the impact – produces the Before–After (BA) design (Figure 1). This compares the system's state before and after the impact, attempting to minimize bias from temporal variability and pre-impact conditions.

Addition of control sites to BA designs results in Before–After Control-Impact (BACI) designs, where the average difference between control and impact sites is compared before and after an impact (Figure 1; Stewart-Oaten et al., 1986; Osenberg et al., 2006). BACI designs use the pre-impact differences between control and impact sites as a null hypothesis for post-impact differences that would exist if the focal impact was absent – avoiding bias from a lack of a control (Thiault, Kernaléguen, Osenberg, & Claudet, 2016). Problems with site-specific temporal variation in BACI designs can be addressed by sampling control and impact sites simultaneously, several times before and after the impact (Before-After Control-Impact Paired-Series (BACIPS) design; Stewart-Oaten & Bence, 2001).

Random allocation of sites to control and impact groups represents the third major aspect of study design. Control-Impact (CI) designs, analogous with Space-For-Time Substitutions (De Palma et al., 2018; França et al., 2016) or Intervention Versus Reference Site designs (Stewart-Oaten & Bence, 2001), compare non-randomly allocated control and impact sites after the impact (Figure 1). However, this non-random allocation can violate the assumption that the only differences between control and impact sites are due to the focal impact, leading to biased results (Damgaard, 2019; De Palma et al., 2018; Larsen, Meng, & Kendall, 2019; Figure 1). Randomized Controlled Trials (RCTs) minimize this bias by randomly allocating experimental units to impact and control groups (Figure 1). This reduces the need to sample before and after the impact to account for any initial differences (i.e. BACI design) if sufficient numbers of sites and points in time are sampled (De Palma et al., 2018; Larsen et al., 2019).

Despite the development of robust approaches to quantifying impacts, greater usage of less robust designs persists. Three systematic maps on the biodiversity impacts of different threats and interventions found that a low proportion of studies used BACI (6%–29%) and BA designs (3%–37%), but many more used CI designs (48%–89%; Bernes et al., 2015, 2017; Papathanasopoulou, Queirós, Beaumont, Hooper, & Nunes, 2016).

The greater prevalence of CI designs in the ecological literature probably reflects that they can be easier to implement than more complex study designs. For example, RCTs are widely used in fields, such as medicine, where random allocation of small-scale experimental units to impact and control groups is possible (Downs & Black, 1998; Tugwell & Haynes, 2006). However, RCTs often cannot be used in ecology because true randomization of experimental units is more difficult with large-scale sites (e.g. protected areas) compared to smaller, more readily-available plots (Larsen et al., 2019; Stewart-Oaten & Bence, 2001). Therefore, ecologists tend to use pseudo-experimental designs lacking randomization, such as BA, CI and BACI designs (Figure 1; De Palma et al., 2018). Nevertheless, constraints due to cost, logistics and project duration often prevent the implementation of complex BACI and even simpler BA designs because of the need to revisit sites pre- and post-impact (França et al., 2016, Osenberg, Shima, Miller, & Stier, 2011; Figure 1).

The disparities between the robustness of study designs and their usage is concerning as many studies may be making misleading inferences about anthropogenic impacts. Some empirical comparisons of the consequences of using BACI, BA and CI designs have been undertaken (Osenberg et al., 2011; França et al., 2016; Mahlum et al., 2018; Smokorowski & Randall, 2017). However, we are yet to

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

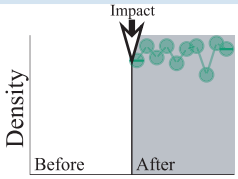
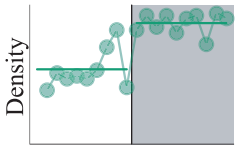
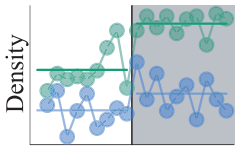
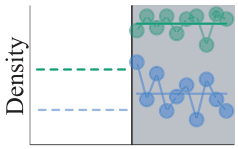
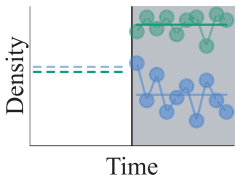
Design	Sampling regime	Relative cost	Relative difficulty in ecology	Suitability	Ecological examples of use
After		Very low	Very low	Most systems Where control unfeasible Unpredictable impacts	Pond creation
Before-After (BA)		Moderate	Moderate	Predictable impacts Where control unfeasible Availability of pre-impact data	Wildlife tunnels under roads
Before-After Control-Impact (BACI) (BARI, MBACI, BACIPS)		High	High	Predictable impacts Appropriate control Availability of pre-impact data	MPA effectiveness, renewable energy infrastructure
Control-Impact (CI) (Space-for-Time, Impact versus Reference Sites)		Low	Moderate	Unpredictable impacts Large-scale replicates that cannot be truly randomised	Oil spill or other pollution event
Randomised Controlled Trial (RCT)		Low	Very high	Unpredictable impacts Small-scale replicates appropriate for randomisation	Peatland restoration, field margins

FIGURE 1 Comparison of the key features of study designs. Graphs show how designs sample from impact (green points) and control (blue points) sites over time, before and after an impact (white vs. grey areas, respectively). Solid horizontal lines show the average density of sites measured to calculate each design's effect size estimate. Dashed horizontal lines for CI and RCTs represent the pre-impact differences between the mean densities of control and impact sites, which can cause bias – note less difference for RCTs (with high sample size) versus CI. Many design variants exist – e.g. MBACI for BACI with multiple sites, R for Reference in BARI (Webb et al., 2012)

understand how inaccurate simpler designs are relative to complex ones, or the influence of sample size on these patterns (e.g. are simpler designs with large sample sizes equivalent to more complex designs with smaller sample sizes?). A quantitative comparison of the accuracy of different designs and their sample size would help us better understand these issues.

To address this knowledge gap, we simulate a hypothetical population's response to an impact, and compare how accurately different study designs estimate that response. We use empirically derived parameter estimates from 47 ecological datasets to generate realistic control and impact data, before and after an impact. BACI, RCT, BA CI and After designs are then used to sample from the simulated data with various levels of spatial replication (control and impact sites). We compare the accuracy of each design by their ability first to predict the correct direction of the response, and second to estimate the response to within a given percentage. Our goal is to inform the development of a quantitative scale of the comparative accuracy of different designs. Such a scale would have utility for future monitoring of anthropogenic impacts, as well as assessing the quality of ecological studies used to inform policy and practice.

2 | MATERIALS AND METHODS

We simulated a hypothetical population with true density λ that varied over T time steps before and after a chronic impact occurred (Figure 2). For example, if $T = 10$, then time steps 1–10 were classified as the 'before period' (i.e. before the impact occurred) and time steps 11–20 were classified as the 'after period' (Figure 2). The true density was monitored in sites where the impact occurred ('impact sites') and where the impact was absent ('control sites').

We set the mean true density to 50 and randomly sampled T values from a Poisson distribution ($\lambda = 50$) to vary the true density over T time steps in the before period for control and impact sites. These T values defined the true density in each time step

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

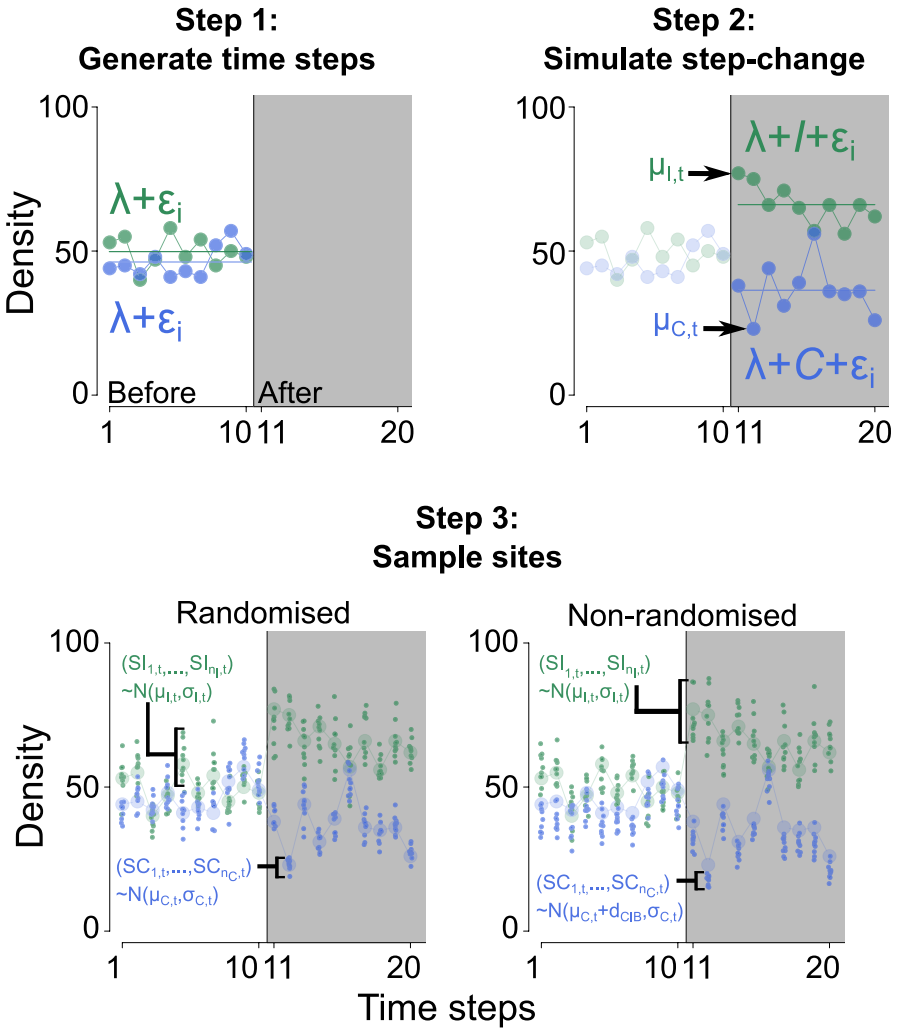


FIGURE 2 An overview of our simulation. Step 1 shows true densities of control and impact sites generated in the before period (white area). Step 2 shows true densities of control and impact sites generated in the after period (grey area) to reflect a step-change response (using *I* and *C*); the true density in each time step (t) is shown ($\mu_{I,t}$, impact: green; and $\mu_{C,t}$, control: blue). Step 3 shows how control and impact sites (*SI* and *SC*) are sampled (n_I and n_C = 10) for both randomized and non-randomized designs

before the impact occurred (e.g. $\mu_{I,t}$ for impact sites in the *t*th time step). To simulate a step-change response at both control and impact sites after the impact occurred (Figure 2), we sampled from a different Poisson distribution with λ adjusted by an empirically derived amount *I* (λ + *I*; Figure 2; Table 1) for impact sites and an empirically derived amount *C* for control sites (λ + *C*; Figure 2; Table 1). *I* and *C* were varied using empirical estimates of the proportional change in control and impact sites in the before period versus the after period, p_I and p_C, respectively, sampled from 47 ecological datasets ($I = \lambda \cdot (p_I - 1)$;$C = \lambda \cdot (p_C - 1)$; Table 1). A list of the published datasets used to calculate empirical estimates in this study are provided in the Data sources section (Burge et al., 2017; Dietl & Durham, 2016; Moland et al., 2013; Sepúlveda & Valdivia, 2016; Williams, Miller, Bright, & Cameron, 2014; see also Appendix S1).

While we focus on a step-change response in our simulation, temporal biodiversity dynamics following disturbances or interventions can follow different trajectories (Di Fonzo, Collen, & Mace, 2013; Thiault et al., 2016). However, to simplify the simulation as much as possible, particularly in terms of computational demands, using a step-change response was most appropriate to test the relative accuracy of each design.

Using the simulated data for before and after periods, we sampled various numbers of impact (n_I) and control (n_C) sites (these could also be plots or transects; Figure 2). For RCTs that use random allocation of sites to control and impact groups, we randomly sampled sites from two normal distributions for each time step: one with a mean, $\mu_{I,t}$, for impact sites and one with a mean, $\mu_{C,t}$, for control sites (Figure 2). The number of sites sampled was the same for all time steps. The standard deviation of each normal distribution represented the variation amongst sites and was calculated by multiplying the mean by the coefficient of variation (e.g. control sites: $\sigma_{C,t} = \mu_{C,t} \cdot CV_S$; impact sites: $\sigma_{I,t} = \mu_{I,t} \cdot CV_S$; Table 1). We varied CV_S by randomly drawing values from a truncated normal distribution: $N(\mu = 0.1, \sigma = 0.05, \min = 0, \max = 0.2)$.

To account for non-random allocation of sites to control and impact groups in BACI, BA, CI and After designs, we repeated the same approach but with one important modification. We adjusted the true density of control sites in every time step, $\mu_{C,t}$, by an empirically derived amount, d_{CIB} ($\mu_{C,t} + d_{CIB}$; Figure 2; Table 1). To vary d_{CIB}, we used empirical estimates of the proportional difference between control and impact sites in the before period, p_{CIB}, sampled from 47 ecological datasets ($d_{CIB} = \lambda \cdot (p_{CIB} - 1)$; Table 1; Appendix S1). This simulated difference between control and impact sites

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Definitions and summary statistics for all simulation parameters (termed 'Sim.') and empirically-derived parameters (termed 'Emp.'; Appendix S1)

Parameter	Definition	Source	Equation	Mean	*SD*	Min	Max
p_C	Change in control between before and after periods	Emp.	$p_C = \frac{\lvert\overline{x_{AC}}\rvert}{\lvert\overline{x_{BC}}\rvert}$	0.918	0.181	0.605	1.31
p_I	Change in impact between before and after periods	Emp.	$p_I = \frac{\lvert\overline{x_{AI}}\rvert}{\lvert\overline{x_{BI}}\rvert}$	0.967	0.230	0.579	1.46
p_{CIB}	Average value of control sites as a proportion of the average value of impact sites in the before period	Emp.	$p_{CIB} = \frac{\lvert\overline{x_{BC}}\rvert}{\lvert\overline{x_{BI}}\rvert}$	1.13	0.306	0.654	1.89
I	True change in impact sites from before to after impact	Emp.	$I = \lambda \cdot (p_I - 1)$	−1.65	11.5	−21.1	23.2
C	True change in control sites from before period to after period	Emp.	$C = \lambda \cdot (p_C - 1)$	−4.10	9.05	−19.8	15.4
d_{CIB}	Difference between true densities of control and impact sites in before period	Emp.	$d_{CIB} = \lambda \cdot (p_{CIB} - 1)$	6.60	15.3	−17.3	44.5
λ	True density across all time steps	Sim.	λ = 50	—	—	50	50
T	Total number of time steps simulated	Sim.	T = {2,4,6,8,10}	—	—	2	10
n_T	Number of time steps sampled in each period	Sim.	$n_T = T$	—	—	2	10
$\mu_{I,t}$	True density in impact sites in time step t	Sim.	Before: $\mu_{I,t} \sim \text{Poisson}(\lambda)$ After: $\mu_{I,t} \sim \text{Poisson}(\lambda + I)$	—	—	—	—
$\sigma_{I,t}$	Standard deviation of impact sites in time step t	Sim.	$\sigma_{I,t} = CV_S \cdot \mu_{I,t}$	—	—	—	—
$\mu_{C,t}$	True density in control sites in time step t	Sim.	Before: $\mu_{C,t} \sim \text{Poisson}(\lambda)$ After: $\mu_{C,t} \sim \text{Poisson}(\lambda + C)$	—	—	—	—
$\sigma_{C,t}$	Standard deviation of control sites in time step t	Sim.	$\sigma_{C,t} = CV_S \cdot \mu_{C,t}$	—	—	—	—
CV_S	Coefficient of variation (variation amongst sites)	Sim.	$CV_s \sim N(\mu, \sigma, \text{min}, \text{max})$	0.10	0.05	0.00	0.20
$SI_{n,t}$	nth impact site sampled in time step t	Sim.	$(SI_{1,t},\ldots,SI_{nI,t}) \sim N(\mu_{I,t}, \sigma_{I,t})$	—	—	—	—
$SC_{n,t}$	nth control site sampled in time step t	Sim.	Randomized: $(SC_{1,t},\ldots, SC_{nc,t}) \sim N(\mu_{C,t}, \sigma_{C,t})$ Non-randomized: $(SC_{1,t},\ldots, SC_{nc,t}) \sim N(\mu_{C,t} + d_{CIB}, \sigma_{C,t})$	—	—	—	—
n_I	Number of impact sites sampled	Sim.	n_I = {1,5,10,25,50}	—	—	1	50
n_C	Number of control sites sampled	Sim.	n_C = {1,5,10,25,50}	—	—	1	50

Note: Equations show how each parameter was calculated. For empirically-derived parameters, $\bar{x}$ refers to the average of sampled sites taken from 47 ecological datasets (e.g. $\overline{x_{AC}}$ refers to the average of all control sites in the after period; Appendix S1).

accounted for different levels of site selection bias in non-randomized designs, including situations where little or no bias may be present (e.g. $d_{CIB} \approx 0$).

We calculated effect size estimates for each design by first finding the mean density of sampled sites across all time steps for control and impact groups in the before period ($\text{Before}_{\text{Impact}}$, $\text{Before}_{\text{Control}}$) and the after period ($\text{After}_{\text{Impact}}$, $\text{After}_{\text{Control}}$). We assumed that sampling occurred in all time steps ($n_T = T$) in both periods. We did this as the investigator may wish to only estimate the effect over a certain timescale (which will be context-specific) and we lacked the computational capabilities to simulate all possible sampling permutations, using fewer than the full number of time steps (e.g. sampling in certain intervals or continuous periods of time; Wauchope, Johnston, Amano, & Sutherland, 2019).

Effect size estimates were calculated using these mean densities, as appropriate for each study design (Table 2). For example, RCT effect sizes were found by subtracting $\text{After}_{\text{Control}}$ from $\text{After}_{\text{Impact}}$, while BA effect sizes were found by subtracting $\text{Before}_{\text{Impact}}$ from $\text{After}_{\text{Impact}}$ (Table 2). The exception was the After design, for which we found the mean of sampled sites in the first time step and subtracted this from the mean of sampled sites in the final time step of the after period (Table 2). We defined the true effect as the change in the mean of true densities of impact sites between the before and after periods minus the equivalent change in the mean of true densities of control sites (Table 2). As discussed previously, we did this because we wanted to compare each design's relative accuracy at estimating the true effect over the number of time steps simulated.

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Equations showing effect size estimate, variance and error calculation for each study design using mean densities of control or impact sites in each period (e.g. $\text{After}_{\text{Impact}}$ refers to the mean of sampled impact sites across all time steps in the after period)

	Equation		
	Effect size estimate	**Pooled variance (s_p^2)**	**Error**
RCT	$\text{After}_{\text{Impact}} - \text{After}_{\text{Control}}$	$\frac{(n_{AI}-1)s_{AI}^2+(n_{AC}-1)s_{AC}^2}{n_{AI}+n_{AC}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{AI}}+\frac{s_p^2}{n_{AC}}}$
BACI	$\left(\begin{array}{c}\text{After}_{\text{Impact}}-\\ \text{After}_{\text{Control}}\end{array}\right)-\left(\begin{array}{c}\text{Before}_{\text{Impact}}-\\ \text{Before}_{\text{Control}}\end{array}\right)$	$\frac{(n_{BI}-1)s_{BI}^2+(n_{BC}-1)s_{BC}^2+(n_{AI}-1)s_{AI}^2+(n_{AC}-1)s_{AC}^2}{n_{BI}+n_{BC}+n_{AI}+n_{AC}-4}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{BI}}+\frac{s_p^2}{n_{BC}}+\frac{s_p^2}{n_{AI}}+\frac{s_p^2}{n_{AC}}}$
CI	$\text{After}_{\text{Impact}} - \text{After}_{\text{Control}}$	$\frac{(n_{AI}-1)s_{AI}^2+(n_{AC}-1)s_{AC}^2}{n_{AI}+n_{AC}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{AI}}+\frac{s_p^2}{n_{AC}}}$
BA	$\text{After}_{\text{Impact}} - \text{Before}_{\text{Impact}}$	$\frac{(n_{BI}-1)s_{BI}^2+(n_{AI}-1)s_{AI}^2}{n_{BI}+n_{AI}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{AI}}+\frac{s_p^2}{n_{BI}}}$
After	$\text{After}_{\text{Impact},\,t=T} - \text{After}_{\text{Impact},\,t=1}$	$\frac{(n_{I,t=T}-1)s_{I,t=T}^2+(n_{I,t=1}-1)s_{I,t=1}^2}{n_{I,t=T}+n_{I,t=1}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{I,t=T}}+\frac{s_p^2}{n_{I,t=1}}}$
True effect	$\left(\frac{\overline{(\mu_{I,\text{After},t=1},\cdots,\mu_{I,\text{After},t=T})}-}{(\mu_{I,\text{Before},t=1},\cdots,\mu_{I,\text{Before},t=T})}\right)-\left(\frac{\overline{(\mu_{C,\text{After},t=1},\cdots,\mu_{C,\text{After},t=T})}-}{(\mu_{C,\text{Before},t=1},\cdots,\mu_{C,\text{Before},t=T})}\right)$		

Note: For the After design, the effect size was calculated by finding the difference between the final time step ($t = T$) and the first time step of the after period (1). n and s^2 refer to the number of sites and variance in that period (e.g. n_{AI} and s_{AI}^2 refer to the number of impact sites and variance in the After period).

We ran the simulation under 1,000 different scenarios, varying: (a) the true change in control sites (C); (b) the true change in impact sites (I); (c) the mean difference between control and impact sites in the before period (d_{CIB}); and (d) the variation between sites (CV_S). For each simulation scenario, we varied the number of time steps simulated (T = 2, 4, 6, 8 or 10), as well as the number of impact sites (n_I = 1, 5, 10, 25, 50) and control sites (n_C = 2, 5, 10, 25, 50) sampled independently to use every possible pairwise combination – a total of 125 combinations. Overall, we simulated 1,000 scenarios with 125 different sampling combinations in each, repeating each scenario 1,000 times ($1{,}000 \times (1{,}000 \times 125) = 1.25 \times 10^8$ runs).

Effect size estimates for each design were used to investigate their relative accuracy. We calculated 95% Confidence Intervals (CIs) using the pooled variance and associated error for each effect size estimate (Table 2). We used these 95% CIs to estimate the percentage of simulations repetitions where: (a) the true effect fell within the 95% CIs of effect size estimates (coverage probability); (b) the correct direction was detected (95% CIs entirely above or below zero); (c) the estimated effect size under- or overestimated the true effect (95% CIs entirely above or below the true effect). We also investigated the percentage of simulation repetitions in which each design's effect size estimate: (a) had the same direction as the true effect; and (b) was both within a given percentage of the true effect *and* of the same direction. We believe these two measures capture the major aspects of accuracy and precision that are desirable in a study design.

We calculated all these measures for all possible pairwise combinations of control and impact sites (e.g. two control and two impact sites, two control and five impact sites etc.). We set five thresholds to measure the percentage of times an effect size estimate was within a certain percentage of the true effect: 10%, 20%, 30%, 40% and 50%. We also explored how varying the terms C and d_{CIB} (controlling bias in non-BACI designs) for three levels of magnitude (no bias: 1; low bias: 0.9 or 1.1; high bias: 0.7 or 1.3) affected this percentage (Figures S7 and S8).

We used Generalized Linear Models with a beta error distribution to determine the relationship between the performance of each design (the response variable; see below) and two explanatory variables (number of control sites and the number of impact sites). For the response variable, we used the proportion of simulation repetitions where the effect size estimate was within ±30% of the true effect and had the correct direction. We only considered results for an accuracy threshold of ±30% as this was deemed a reasonable level of accuracy and we wanted to simplify the interpretation of our results as much as possible. We also present results for other accuracy thresholds (±10% and ±50%) in Appendix S3.

Based on graphical observations of the relationship between the response and explanatory variables (Figure 5), we included impact and control sites as log transformed explanatory variables for models of BACI and RCT designs and tested models with and without an interaction term between these variables (Appendix S4). The BACI model with the interaction term had the lowest AIC by more than 2 units and was chosen as the best model. The RCT model without the interaction term was only lower by 1.8 units, but was chosen as it was more parsimonious (Appendix S4). As the performance of BA, CI and After designs did not vary with the number of impact or control sites, we did not create any models for these designs (Figure 5). We calculated quasi-R^2 values (Appendix S4) to test model performance using the equation:

$$\text{quasi}R^2 = 1 - \frac{\text{deviance}}{\text{null deviance}} \quad (1)$$

Both models were only slightly over-dispersed (RCT model: θ = 1.19; BACI model: θ = 1.25) and Pearson's χ^2 residuals were non-significant ($p > .05$), suggesting no significant patterns remained in

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the residuals. There were also no observable patterns between residuals and explanatory variables or fitted values.

For BACI and RCT designs, we converted estimated coefficients (β) from log odds (Appendix S4) to proportions to create an 'accuracy weight' equation for each design (Equations 2 and 3). We found the accuracy weights for BA, CI and After designs by simply taking the mean proportion of simulation repetitions where the effect size estimate was within ±30% of the true effect and had the correct direction across all combinations of impact and control sites. These accuracy weights are on a continuous scale between a minimum of 0 (lowest accuracy) and a maximum of 1 (highest accuracy – see Results, Discussion and Appendix S2 as to how to apply these weights):

$$\text{BACI accuracy weight} = \frac{1}{1+e^{-\left(\begin{array}{c}\beta_{Int.}+\beta_{n_I}\cdot\ln(n_I)+\beta_{n_C}\cdot\ln(n_C)\\+\beta_{n_In_c}\cdot\ln(n_I)\cdot\ln(n_C)\end{array}\right)}} \quad (2)$$

$$\text{RCT accuracy weight} = \frac{1}{1+e^{-\left(\beta_{Int.}+\beta_{n_I}\cdot\ln(n_I)+\beta_{n_C}\cdot\ln(n_C)\right)}}, \quad (3)$$

where $\beta_{Int.}$ = Intercept coefficient, β_{n_I} = Impact sites coefficient, β_{n_C} = Control sites coefficient and $\beta_{n_In_c}$ = interaction coefficient between impact and control sites.

We applied our accuracy weights to three recent ecological meta-analyses: Bernes et al. (2018) on the effects of ungulate herbivory on vegetation and invertebrates; Eales et al. (2018) on the effects of prescribed burning on forest biodiversity; and Sandström et al. (2019) on the impacts of dead wood manipulation on forest biodiversity. We found these meta-analyses by searching the Environmental Evidence Journal and the Journal of Applied Ecology using the search terms: 'meta analysis' OR 'meta-analysis' and reviewing studies published since 2018. Only the three previously mentioned meta-analyses contained a sufficient range of study designs (Appendix S2) and readily available associated data on study design, replicates and effect sizes. We repeated analyses using random effects models following the authors' methodology (e.g. including random factors such as Site IDs) using the metaphor package (Viechtbauer, 2010; see Appendix S2). We were able to replicate 128 out of 130 summary effect sizes (comparisons) using the authors' methodology and inverse-variance weighting, which we repeated with our accuracy weights. Two summary effect sizes could not be replicated from Bernes et al. (2018) due to lack of data labelling. We wanted to test how our weights altered the conclusions of meta-analyses that used studies with a mixture of different study designs. Therefore, we only present results for 96 comparisons that used studies with at least one type of design. The mean number of studies of each design was as follows: 9.0 BACI, 6.0 BA, 5.0 CI (see Appendix S2 for a breakdown of studies for each summary effect size).

We used R statistical software version 3.5.1 (R Core Team, 2018) with the doParallel package (Microsoft Corporation & Weston, 2017) to increase computational performance. We provide data to repeat all analyses on Zenodo (Christie et al., 2019).

3 | RESULTS

There was large variation in the performance of designs in accurately estimating the true effect. As overall patterns were similar across simulations with different time steps (Figures S1–S3), we present results when six time steps were simulated in both the before and after periods.

BACI designs performed best at correctly identifying the direction of the true effect (≥94.1% of simulation repetitions; Figure 3a), followed by RCTs (≥90.8% of the time). Both BACI and RCTs far outperformed CI, BA, and particularly After designs – BA designs

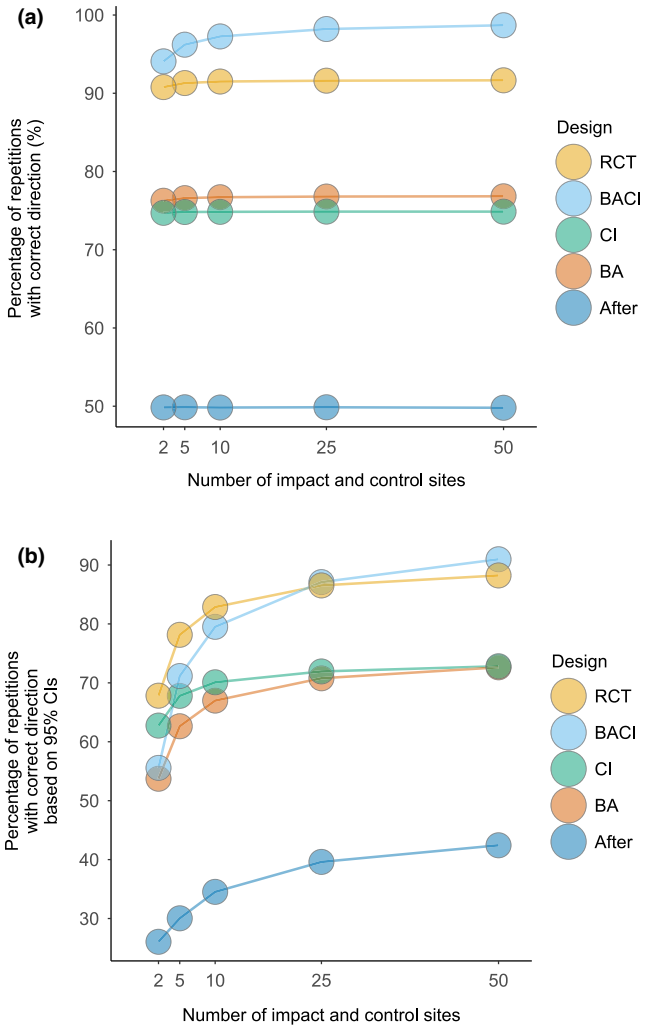


FIGURE 3 Performance of designs in correctly predicting the direction of the true effect for multiple levels of spatial replication with equal numbers of control and impact sites (see Figures S1 and S2 for other combinations of sites). Part (a) measures this in terms of whether the effect size estimate was positive or negative, while part (b) considers whether the 95% CIs of this estimate correctly fell entirely above or below zero. See Figure 1 for the definition of each design

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

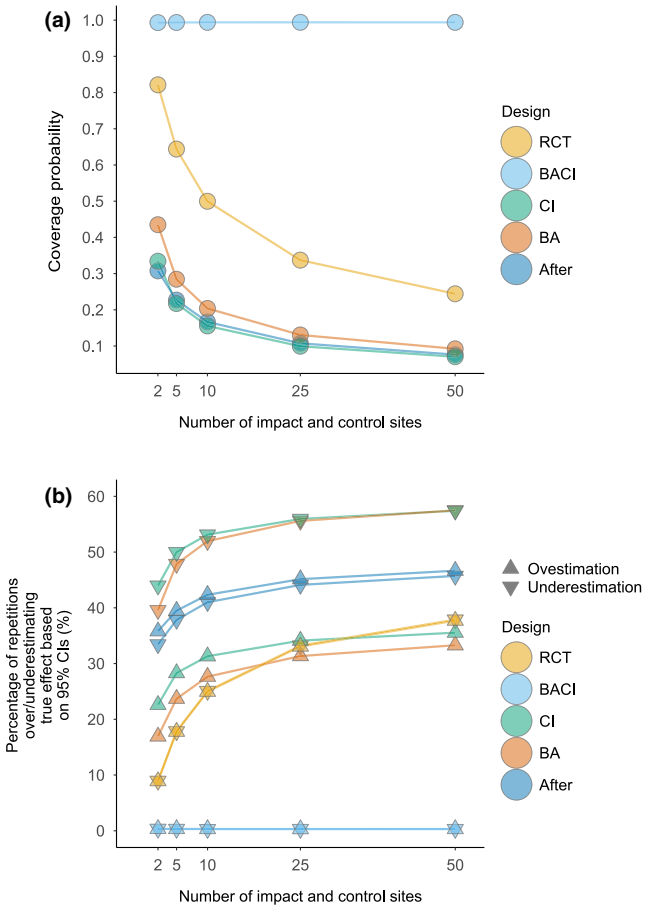


FIGURE 4 Percentage of simulation repetitions in which the 95% CIs of effect size estimates contained the true effect (coverage probability – (a)) or were either greater than or less than the true effect (overestimate vs. underestimate – (b)). In (b), underestimates are shown by downward triangles, whilst overestimates are shown by upward triangles. This is shown for multiple levels of spatial replication with equal numbers of control and impact sites (see Figures S4 and S5 for other combinations of sites). See Figure 1 for the definition of each design

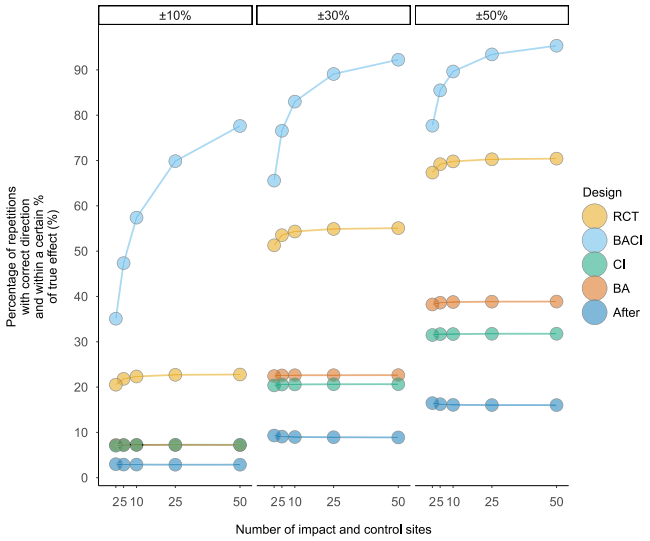


FIGURE 5 Performance of designs measured by percentage of simulation repetitions in which a design's effect size estimate was both within ±10, 30 or 50% of the true effect and had the correct direction. This is shown for multiple levels of spatial replication with equal numbers of control and impact sites (see Figure S6 for other combinations of sites). See Figure 1 for the definition of each design

slightly outperformed CI designs (approximately 76.3% vs. 74.7%) and both BA and CI designs strongly outperformed After designs (approximately 49.8%; Figure 3a). Unlike BACI designs, non-BACI designs showed negligible improvements in performance with increasing replication (increases from two control and two impact sites to 50 control and 50 impact sites: BACI = +4.6%; After = +0.0%; BA = +0.5%; CI = +0.2%; RCT = +1.0%; Figure 3a).

Taking account of the uncertainty around these effect size estimates (95% CIs) gave different results – where overlap with zero was classed as non-significant and non-overlap as either positive or negative (Figure 3b). With this measure, RCTs were most likely to correctly predict the direction of the true effect with two impact and control sites, followed by CI, BACI, BA and After designs in decreasing order of performance (Figure 3b). BACI designs showed the greatest proportional improvement in this measure, outperforming RCTs at sample sizes above 25 impact and control sites (Figure 3b). BA designs increased proportionally more than CI designs, reaching similar levels of performance above 25 impact and control sites (Figure 3b). BACI designs were also likely to produce non-significant effect sizes (overlap of 95% CIs with zero in ~45% of repetitions for two impact and control sites; Figure S3) at low sample sizes, but were extremely unlikely to produce significant effect sizes that had the wrong direction (~1% of repetitions for 2 impact and control sites; Figure S3).

If we consider the coverage probabilities of each design (i.e. proportion of times the true effect fell within the 95% CIs of effect size estimates), BACI designs substantially outperformed other designs (Figure 4a). The true effect fell within the 95% CIs of BACI effect size estimate approximately 99% of the time, with negligible change in increasing replication (Figure 4a). The coverage probabilities for other designs (RCT, BA, CI, After) declined asymptotically with increasing replication as 95% CIs narrowed (Figure 4a).

We also examined the tendency for designs to underestimate or overestimate the true effect (i.e. when 95% CIs did not overlap with true effect; Figure 4b). BACI designs rarely gave under- or overestimates of the true effect (1% of repetitions), while both BA and CI designs were approximately twice as likely to give underestimates than overestimates (Figure 4b). RCT designs and (to a lesser extent) After designs were equally as likely to yield underestimates as overestimates. With increasing replication, all non-BACI designs were increasingly likely to under- or overestimate the true effect, although this relationship was asymptotic – this probability increased at a higher rate for RCTs than other non-BACI designs (Figure 4b).

Consistent patterns were also found when considering the percentage of repetitions for which the effect size estimate was both within a certain percentage of the true effect and had the correct direction (Figure 5). First, RCT and BACI designs still

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Comparison of outcomes for 96 summary effect sizes obtained using the accuracy weights proposed by this study versus conventional inverse-variance weighting

		Accuracy weight		
Weighting method	**Outcome**	**+**	**Non-significant**	**–**
Inverse-variance weight	+	9 (9%)	8 (8%)	0
	Non-significant	1 (1%)	53 (55%)	2 (2%)
	–	0	11 (12%)	12 (13%)

Note: Summary effect sizes were extracted from three separate meta-analyses (see Materials and Methods). Cells show the proportion of effect sizes that were significantly positive, significantly negative or non-significant for both weighting systems.

far outperformed CI, BA or After designs (for ±30% accuracy threshold: BACI ≥ 65.6%, RCT ≥ 51.3%, BA ≥ 22.4%, CI ≥ 20.4%, After ≥ 9.3%; Figure 5). Second, BA designs appeared to perform slightly better than CI designs, especially as the accuracy threshold rose from ±10% to −50% (from ~2% higher to ~7% higher performance; Figure 5). Similarly, both BA and CI designs performed relatively better compared to After designs with an increasing accuracy threshold (Figure 5).

Third, BACI performance increased to a much greater extent with increasing replication than for other designs (Figure 5). For the ±30% accuracy threshold, increasing replication from two control and impact sites to 50 control and impact sites resulted in an increase of 26.7% for BACI compared to +3.8% for RCT, +0.2% for BA, +0.3% for CI and −0.4% for After (Figure 5). For BACI designs, increasing replication moderately in both control and impact sites resulted in greater performance than only increasing replication in just one type of site (±30% threshold: 76.6% at two impact and two control sites vs. 72.8% at two impact and 50 control sites; Figure 4; Figure S6).

We also considered how varying the simulation parameters C and d_{CIB} affected our results (Figures S7 and S8). Increasing the change in control (C) reduced the performance of BA designs substantially (Figure S7), while increasing the initial mean differences between impact and control groups in the before period (d_{CIB}) reduced the performance of CI designs substantially (Figure S8).

We used Generalized Linear Models to examine the factors that determined the performance of each design at estimating the direction and magnitude of the true effect to within ±30% (using data from Figure 5 and Figure S6). For RCT designs, there was little difference in the importance of control versus impact sites in predicting performance, while control sites seemed to have greater importance in BACI designs. High Pseudo-R^2 values showed that our models explained far greater levels of variation in the data than null models (see Appendix S4).

Weights for studies in meta-analyses can be calculated from these relationships of performance with sample size, which we term 'accuracy weights'. This requires information about a study's design and the number of independent control and impact units used (see Appendix S4). For example, França et al. (2016) used a BACI design, 29 impact and five control units and thus receives an accuracy weight of as follows:

$$\frac{1}{1+e^{-\left(\begin{array}{c}0.661+0.0153\cdot\ln(29)+0.0647\cdot\ln(5)\\+0.111\cdot\ln(29)\cdot\ln(5)\end{array}\right)}}=0.805$$

(see Appendix S4; Equation 2). More examples of calculating weights for studies are shown in Appendix S2.

We applied our accuracy weights to meta-analyses (Appendix S2) and found that they gave broadly similar results to conventional (inverse-variance) weighting (for 77% of comparisons). However, there was a tendency for our weights to alter the outcome to non-significant (12% from negative to non-significant and 8% from positive to non-significant; Table 3). A small proportion changed from non-significant to significantly positive (1%) or significantly negative (2%). No outcomes of summary effect sizes changed from positive to negative or vice versa (Table 3).

4 | DISCUSSION

Using this simulation, we have demonstrated that BACI and RCT designs are far more accurate than BA, CI and After designs. When estimating the true effect to within ±30% and correctly identifying its direction, BACI designs performed: 1.3–1.8 times better than RCTs; 2.9–4.2 times better than BA; 3.2–4.6 times better than CI; and 7.1–10.1 times better than After designs (depending on sample size). This is because increasing sample size tends to only increase precision in non-BACI designs around a biased estimate of the true effect.

This bias is generated by violating the assumptions underpinning these non-BACI designs (De Palma et al., 2018). BA designs assume there is no average change in control group mean before versus after the intervention, whilst CI designs assume the only differences that exist between control and impact sites are due to the focal impact (C and d_{CIB} in Materials and Methods; Figures S7 and S8). RCTs also suffer from bias because they only minimize initial differences between impact and control group means (in the before period), which can be increased by spatiotemporal variation (De Palma et al., 2018). Blocking, pairing or matching sites in CI designs or including a proxy variable in statistical analysis could theoretically account for some of these biases, but they cannot guarantee lower levels of bias. BACI designs better account for these initial differences, effectively removing this bias, and therefore have higher accuracy – they can also deal with spatiotemporal variation through repeated sampling through time (Thiault et al., 2016).

The fact that increasing the sample size (precision) of non-BACI designs reduced the coverage probability (probability that the true effect fell within the 95% CIs of an effect size estimate; Figure 4a)

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

supports this conclusion as 95% CIs converged on biased estimates. The coverage probability for BACI designs remained almost constant because the effect size estimates converged on the true effect – greater precision translated into greater accuracy for BACI designs. Greater accuracy is therefore best achieved using more robust designs that remove biases, such as BACI, with greater sample sizes.

We, nevertheless, note that BACI designs are known to suffer from noise (Osenberg et al., 2006), which was demonstrated by their tendency to classify the direction of the true effect as non-significant at low samples sizes (Figure S3). The low statistical power of BACI designs at small sample sizes is an issue and reinforces the need to ensure sufficient numbers of replicates are used in BACI designs (Osenberg et al., 2006), as well as to consider using Bayesian approaches to interpret effect sizes (Conner, Saunders, Bouwes, & Jordan, 2016). When only considering the direction of effect size estimates, however, BACI and RCT designs gave the correct direction ~20%–30% more often than CI and BA designs, and ~40%–45% more than After designs (Figure 3a).

Our results provide strong evidence that simpler designs (e.g. After, BA and CI) often yield different inferences to BACI designs, as observed empirically by previous studies (França et al., 2016; Mahlum et al., 2018; Osenberg et al., 2011; Smokorowski & Randall, 2017). We also found that BA and CI designs were more prone to underestimation than overestimation (Figure 4b), which is consistent with results from França et al. (2016) that showed a CI design underestimated the impacts of logging relative to a BACI design. Therefore, we argue that studies using After, BA and CI designs risk presenting misleading conclusions on the impact of threats and interventions. To our knowledge, this simulation is not only the first quantitative comparison to demonstrate this, but also to show *how* inaccurate non-BACI designs may be, on average in ecology, under varying levels of spatial replication.

We have confidence in these conclusions as we used empirically derived parameter estimates from 47 ecological datasets to quantify the likelihood and magnitude of the biases that affect study designs in ecology (d_{CIB} and C; Materials and Methods; Appendix S1). The context-dependency of our results, linked to how the likelihood and magnitude of biases varies across different fields of ecology, could be investigated using our R code if sufficient empirical data are available to characterize major parameters (d_{CIB} and C; Figures S7 and S8) in different contexts. Future work could explore the effects of different types of trends and lag periods on the relative performance of designs, since previous literature has often assumed there is no overall pre-impact trend – only fluctuations around a baseline average (Lettenmaier et al., 1978; Thiault et al., 2016). Nevertheless, our results provide strong evidence that, generally in ecology, we should invest in implementing more robust designs whenever possible – investing effort into using simpler designs with greater sample sizes is simply inefficient.

Although we strongly advocate for greater investment in more robust designs, we also realize there is a trade-off between the greater accuracy of robust designs and greater logistical ease of simpler designs. While we generate more studies with simpler designs more easily, their probable low accuracy means that we may use misleading evidence to inform policy and practice. We, nevertheless, argue that situations still arise where investigators can use robust designs and yet fail to; promoting greater awareness of more robust designs and opportunities for their usage is important. For example, BACI design usage should be encouraged whenever prior knowledge exists in the timing of an impact or where suitable pre-impact data is available retrospectively (e.g. infrastructure projects, Protected Area designation). We also recognize the expensive nature of BACI designs, due to the need to revisit study sites before and after the impact, often hampers their implementation (De Palma et al., 2018). This means BACI designs can be impossible to use during short-term projects limited by grant or studentship duration. Therefore, we suggest that longer-term funding and stronger research–practice partnerships are required to facilitate the use of BACI designs (De Palma et al., 2018; Osenberg et al., 2011).

Some investigators may also avoid using BACI designs due to concerns over the difficulty in interpreting their results for lay audiences and stakeholders. We suggest that this can be overcome using Bayesian approaches that present easily interpretable probabilities to managers and practitioners (Conner et al., 2016) or new measures for BACI designs that aid ecological interpretation (Chevalier, Russell, & Knape, 2019).

Given the use of simpler designs will probably persist in the near future, we further argue that our results have major implications for decision-making and meta-analysis in ecology. We have proposed a novel weighting system that could help when meta-analyses are faced with studies that vary markedly in their design. Conventional meta-analysis typically uses inverse-variance of studies as weights to attempt to account for study quality (Koricheva & Gurevitch, 2014; Marín-Martínez & Sánchez-Meca, 2010). However, this can greatly reduce the number of suitable primary studies since not all studies report variance (Koricheva & Gurevitch, 2014). Alternative approaches of meta-analysis to tackle poor data reporting, such as non-parametric weighting by sample size, have been proposed (Adams, Gurevitch, & Rosenberg, 1997; Mayerhofer, Kernaghan, & Harper, 2013), but fail to consider wider aspects of study quality such as study design (Spake & Doncaster, 2017). While recent efforts to assessing evidence quantitatively by study design are welcomed (Mupepele & Dormann, 2017; Mupepele, Walsh, Sutherland, & Dormann, 2016; Webb et al., 2012), their weights are relatively simplistic (e.g. simple integer scores or categories) and lack a quantitative or objective grounding.

Our weighting system is informed by the relationships we found between accuracy, precision, study design and sample size, arguably accounting for more aspects of study quality than weighting by sample size or inverse-variance. We have shown how our accuracy weights can be easily used in meta-analyses to give greater influence to studies with more accurate designs (Appendix S2). We have also demonstrated that our weights tend to reduce the number of positive and negative significant results in meta-analyses compared to using inverse-variance weighting (Table 3). We argue that our simulation results imply that inverse-variance weighting may erroneously

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reward studies with non-BACI designs when they have higher precision (lower variance), possibly leading to more significantly positive or negative results. This is problematic because we have shown that increasing precision of non-BACI designs often leads to biased estimates and not greater accuracy. Weighting by a combination of accuracy and precision using our weights seems more sensible given these results.

Although we acknowledge that our weights only consider some aspects of study quality, we believe that they could be modulated using the percentage of criteria met in subject-specific quality checklists to incorporate more context-specific factors (e.g. size of sampling unit, temporal replication and internal validity; Bilotta, Milner, & Boyd, 2014; Mupepele et al., 2016). Adding extra components to the evidence assessment process, however, must be balanced against the effort expended in doing so. Our weights could also assign studies to different accuracy categories (Appendix S2), giving a rapid, easily interpretable way to communicate the robustness of evidence to decision-makers – e.g. in evidence toolkits such as Conservation Evidence (Sutherland et al., 2019). We welcome future research to explore how best to apply our accuracy weights within evidence assessment and decision-making processes.

Overall, we have shown for the first time how much less accurate simpler study designs are compared to more complex ones, generating a new quantitative understanding of the relative accuracy of different designs. Our accuracy weights could also offer a powerful, yet versatile new approach to weighting evidence where studies use a range of different designs, with major implications for the future of meta-analysis and decision-making. We hope our work encourages greater discussion of study design by scientists, managers and policy-makers across ecology and demonstrates the need to tackle the serious consequences of using different designs to make inferences in ecology.

ACKNOWLEDGEMENTS

We thank the following people for providing empirical data used to parameterize simulation: Anna Sher, Ricardo Rocha, Annelies de Backer, Aurora Torres, Carlos Palacín, Juan Carlos Alonso, Barry Baldigo, Brendan Kelaher, Daniel Mateos, Doriane Stagnol, Dominique Davoult, Filipe França, Heather Major, Ian Jones, Jake Bicknell, Jenyffer Vieira, Maria Ruiz-Delgado, Ruben Heleno, Joachim Claudet, Kade Mills, Kevin Stokesbury, Bradley Harris, Mehdi Adjeroud, Michael Craig, Michele Meroni, Norbertas Noreika, Janne Kotiaho, Patrick Edwards, Rafael Barrientos, Carlos Ponce, Carlos Martín, Beatriz Martín, Ricardo Ceia, Roland Pitcher, Sarah Clarke, Oliver Tully, Shailesh Sharma, Just Cebrian, Thomas Stanley, Tyler Eddy, Jonathan Gardner, Anjali Pande, Adrià López-Baucells, Christoph Meyer, Alvaro Antón, Bob McConnaughey, Corrine Watts, David Abecasis, Luciana Cibils, Monica Montefalcone, Teppo Vehanen, Aki Mäki-Petäys, Ari Huusko, Juan Schmitter-Soto, Matt Rinella, Garth Hodgson, Hartwell Welsh, Mikael van Deurs, Mary Donovan, Axel Schwerk, Jill Shaffer, Deborah Buhl, Alberto Velando, Dolores River Restoration Partnership, Javier Pinilla, Andrew Page, Matt Dasey, David Maguire, Jos Barlow, Júlio Louzada, Rachel Buxton, Carley Schacter, Melinda Conners, Koniambo Nickel, Ginger Soproner, CSIRO, Arturo Elosegi, Loreto García-Arberas, Joserra Díez and Ana Rallo. Thanks also to Grania Smith for proofreading drafts. Author funding sources: T.A. was supported by the Grantham Foundation for the Protection of the Environment, Kenneth Miller Trust and Australian Research Council Future Fellowship (FT180100354); W.J.S., P.A.M. and G.E.S. were supported by Arcadia and The David and Claudia Harding Foundation; B.I.S. and A.P.C. were supported by the Natural Environment Research Council via Cambridge Earth System Science NERC DTP (NE/L002507/1).

AUTHORS' CONTRIBUTIONS

A.P.C., W.J.S. and T.A. conceived the ideas and designed simulation; A.P.C. analysed the data and led the writing of the manuscript. All authors critically contributed to drafts as well as giving the final approval for publication. T.A. provided the Japanese abstract.

DATA AVAILABILITY STATEMENT

Simulation R script and empirical data used in simulations are available via Zenodo https://doi.org/10.5281/zenodo.3373462 (Christie et al., 2019). Published datasets from which some of this empirical data was derived are cited in the Data Sources section.

ORCID

Alec P. Christie https://orcid.org/0000-0002-8465-8410
Tatsuya Amano https://orcid.org/0000-0001-6576-3410
Philip A. Martin https://orcid.org/0000-0002-6804-2720
Gorm E. Shackelford https://orcid.org/0000-0003-0949-0934
Benno I. Simmons https://orcid.org/0000-0002-2751-9430
William J. Sutherland https://orcid.org/0000-0002-6498-0437

REFERENCES

Adams, D. C., Gurevitch, J., & Rosenberg, M. S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology*, *78*, 1277–1283. https://doi.org/10.1890/0012-9658(1997)078[1277:RTFMAO]2.0.CO;2

Bernes, C., Bullock, J. M., Jakobsson, S., Rundlöf, M., Verheyen, K., & Lindborg, R. (2017). How are biodiversity and dispersal of species affected by the management of roadsides? *A Systematic Map. Environmental Evidence*, *6*(1), 1–16. https://doi.org/10.1186/s13750-017-0103-1

Bernes, C., Jonsson, B. G., Junninen, K., Lõhmus, A., Macdonald, E., Müller, J., & Sandström, J. (2015). What is the impact of active management on biodiversity in boreal and temperate forests set aside for conservation or restoration? A systematic map. *Environmental Evidence*, *4*(1), 25. https://doi.org/10.1186/s13750-015-0050-7

Bilotta, G. S., Milner, A. M., & Boyd, I. L. (2014). Quality assessment tools for evidence from environmental science. *Environmental Evidence*, *3*(1), 14. https://doi.org/10.1186/2047-2382-3-14

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*(349), 70–79. https://doi.org/10.1080/01621459.1975.10480264

Chevalier, M., Russell, J. C., & Knape, J. (2019). New measures for evaluation of environmental perturbations using Before-After-Control-Impact analyses. *Ecological Applications*, *29*(2), e01838. https://doi.org/10.1002/eap.1838

Christie, Alec P., Amano, Tatsuya, Martin, Philip A., Shackelford, Gorm E., Simmons, Benno I., & Sutherland, William J. (2019). alecchristie888/studydesignsim: First release (Version v1.0.0). [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.3373462

Conner, M. M., Saunders, W. C., Bouwes, N., & Jordan, C. (2016). Evaluating impacts using a BACI design, ratios, and a Bayesian approach with a focus on restoration. *Environmental Monitoring and Assessment*, *188*(10), 555. https://doi.org/10.1007/s10661-016-5526-6

Damgaard, C. (2019). A critique of the space-for-time substitution practice in community ecology. *Trends in Ecology & Evolution*, *34*(5), 416–421. https://doi.org/10.1016/j.tree.2019.01.013

De Palma, A., Sanchez Ortiz, K., Martin, P. A., Chadwick, A., Gilbert, G., Bates, A. E., … Purvis, A. (2018). Challenges with inferring how land-use affects terrestrial biodiversity: Study design, time, space and synthesis. *Next Generation Biomonitoring*, *58*, 163–199. https://doi.org/10.1016/bs.aecr.2017.12.004

Di Fonzo, M., Collen, B., & Mace, G. M. (2013). A new method for identifying rapid decline dynamics in wild vertebrate populations. *Ecology and Evolution*, *3*(7), 2378–2391. https://doi.org/10.1002/ece3.596

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, *52*(6), 377–384. https://doi.org/10.1136/jech.52.6.377

França, F., Louzada, J., Korasaki, V., Griffiths, H., Silveira, J. M., & Barlow, J. (2016). Do space-for-time assessments underestimate the impacts of logging on tropical biodiversity? An Amazonian case study using dung beetles. *Journal of Applied Ecology*, *53*(4), 1098–1105. https://doi.org/10.1111/1365-2664.12657

Hewitt, J. E., Thrush, S. E., & Cummings, V. J. (2001). Assessing environmental impacts: Effects of spatial and temporal variability at likely impact scales. *Ecological Applications*, *11*(5), 1502–1516. https://doi.org/10.1890/1051-0761(2001)011[1502:AEIEOS]2.0.CO;2

Hipel, K. W., Lettenmaier, D. P., & McLeod, A. I. (1978). Assessment of environmental impacts part one: Intervention analysis. *Environmental Management*, *2*(6), 529–535. https://doi.org/10.1007/BF01866711

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Ecology*, *102*, 828–844. https://doi.org/10.1111/1365-2745.12224

Larsen, A. E., Meng, K., & Kendall, B. E. (2019). Causal analysis in control-impact ecological studies with observational data. *Methods in Ecology and Evolution*, *10*(7), 924–934. https://doi.org/10.1111/2041-210X.13190

Lettenmaier, D. P., Hipel, K. W., & McLeod, A. I. (1978). Assessment of environmental impacts part two: Data collection. *Environmental Management*, *2*(6), 537–554. https://doi.org/10.1007/BF01866712

Mahlum, S., Cote, D., Wiersma, Y. F., Pennell, C., & Adams, B. (2018). Does restoration work? It depends on how we measure success. *Restoration Ecology*, *26*(5), 952–963. https://doi.org/10.1111/rec.12649

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*(1), 56–73. https://doi.org/10.1177/0013164409344534

Mayerhofer, M. S., Kernaghan, G., & Harper, K. A. (2013). The effects of fungal root endophytes on plant growth: A meta-analysis. *Mycorrhiza*, *23*, 119–128. https://doi.org/10.1007/s00572-012-0456-9

Microsoft Corporation, & Weston, S. (2017). doParallel: Foreach parallel adaptor for the 'parallel' package. R package version 1.0.11. Retrieved from https://CRAN.R-project.org/package=doParallel

Mupepele, A. C., & Dormann, F. C. (2017). Influence of forest harvest on nitrate concentration in temperate streams—A meta-analysis. *Forests*, *8*(1), 5. https://doi.org/10.3390/f8010005

Mupepele, A. C., Walsh, J. C., Sutherland, W. J., & Dormann, C. F. (2016). An evidence assessment tool for ecosystem services and conservation studies. *Ecological Applications*, *26*(5), 1295–1301. https://doi.org/10.1890/15-0595

Osenberg, C. W., & Schmitt, R. J. (1996). Detecting ecological impacts caused by human activities. In R. J. Schmitt & C. W. Osenberg (Eds.), *Detecting ecological impacts: Concepts and applications in coastal habitats* (pp. 3–16). San Diego, CA: Academic Press.

Osenberg, C. W., Bolker, B. M., White, J. S. S., St Mary, C. M., & Shima, J. S. (2006). Statistical issues and study design in ecological restorations: Lessons learned from marine reserves. In A. D. Falk, M. A. Palmer, & J. B. Zedler (Eds.), *Foundations of restoration ecology* (pp. 280–302). Washington, DC: Island Press.

Osenberg, C. W., Shima, J. S., Miller, S. L., & Stier, A. C. (2011). Assessing effects of marine protected areas: Confounding in space and possible solutions. In J. Claudet (Ed.), *Marine protected areas: A multidisciplinary approach* (pp. 143–167). Cambridge, UK: Cambridge University Press.

Papathanasopoulou, E., Queirós, A. M., Beaumont, N., Hooper, T., & Nunes, J. (2016). What evidence exists on the local impacts of energy systems on marine ecosystem services: A systematic map. *Environmental Evidence*, *5*(1), 1–12. https://doi.org/10.1186/s13750-016-0075-6

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Smokorowski, K. E., & Randall, R. G. (2017). Cautions on using the Before-After-Control-Impact design in environmental effects monitoring programs. *Facets*, *2*(1), 212–232.

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. https://doi.org/10.1016/j.foreco.2017.05.059

Stewart-Oaten, A., & Bence, J. R. (2001). Temporal and spatial variation in environment impact assessment. *Ecological Monographs*, *71*, 305–339.

Stewart-Oaten, A., Murdoch, W. W., & Parker, K. R. (1986). Environmental impact assessment: 'Pseudoreplication' in time? *Ecology*, *67*, 929–940. https://doi.org/10.2307/1939815

Sutherland, W. J., Taylor, N. G., MacFarlane, D., Amano, T., Christie, A. P., Dicks, L. V., … Wordley, C. F. R. (2019). Building a tool to overcome barriers in research-implementation spaces: The conservation evidence database. *Biological Conservation*, *238*, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Thiault, L., Kernaléguen, L., Osenberg, C. W., & Claudet, J. (2016). Progressive-change BACIPS: A flexible approach for environmental impact assessment. *Methods in Ecology and Evolution*, *8*(3), 288–296. https://doi.org/10.1111/2041-210X.12655

Tugwell, B., & Haynes, R. B. (2006). Assessing claims of causation. In B. Tugwell, R. B. Haynes, R. B. Haynes, D. L. Sackett, G. H. Guyatt, & P. Tugwell (Eds.), *Clinical epidemiology: How to do clinical practice research* (pp. 356–387). Philadelphia, PA: The University of Chicago Press.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/.

Wauchope, H. S., Johnston, A., Amano, T., & Sutherland, W. J. (2019). When can we trust population trends? A method for quantifying the effects of sampling interval and duration, *bioRxiv*, 498170.

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Webb, J. A., Nichols, S. J., Norris, R. H., Stewardson, M. J., Wealands, S. R., & Lea, P. (2012). Ecological responses to flow alteration: Assessing causal relationships with eco evidence. *Wetlands*, *32*(2), 203–213. https://doi.org/10.1007/s13157-011-0249-5

Wolfe, D. A., Champ, M. A., Flemer, D. A., & Mearns, A. J. (1987). Long-term biological datasets: Their role in research, monitoring and management of estuarine and coastal marine systems. *Estuaries*, *10*, 181–193. https://doi.org/10.2307/1351847

DATA SOURCES

Bernes, C., Macura, B., Jonsson, B. G., Junninen, K., Müller, J., Sandström, J., … Macdonald, E. (2018). Manipulating ungulate herbivory in temperate and boreal forests: Effects on vegetation and invertebrates. A systematic review. *Environmental Evidence*, *7*(1), 13. https://doi.org/10.1186/s13750-018-0125-3

Burge, O. R., Bodmin, K. A., Clarkson, B. R., Bartlam, S., Watts, C. H., & Tanner, C. C. (2017). Data from: Glyphosate redirects wetland vegetation trajectory following willow invasion. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.1nn35

Dietl, G. P., & Durham, S. R. (2016). Data from: Geohistorical records indicate no impact of the Deepwater Horizon oil spill on oyster body size. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.bc80t

Eales, J., Haddaway, N. R., Bernes, C., Cooke, S. J., Jonsson, B. G., Kouki, J., … Taylor, J. J. (2018). What is the effect of prescribed burning in temperate and boreal forest on biodiversity, beyond pyrophilous and saproxylic species? A systematic review. *Environmental Evidence*, *7*(1), 19. https://doi.org/10.1186/s13750-018-0131-5

Moland, E., Olsen, E. M., Knutsen, H., Garrigou, P., Espeland, S. H., Kleiven, A. R., … Knutsen, J. A. (2013). Lobster and cod benefit from small-scale northern marine protected areas: Inference from an empirical before–after control-impact study. *Proceedings of the Royal Society B*, *280*, 20122679. https://doi.org/10.1098/rspb.2012.2679

Sandström, J., Bernes, C., Junninen, K., Lõhmus, A., Macdonald, E., Müller, J., & Jonsson, B. G. (2019). Impacts of dead wood manipulation on the biodiversity of temperate and boreal forests. A systematic review. *Journal of Applied Ecology*, *56*, 1770–1781. https://doi.org/10.1111/1365-2664.13395

Sepúlveda, R. D., & Valdivia, N. (2016). Localised effects of a mega-disturbance: Spatiotemporal responses of intertidal sandy shore communities to the 2010 Chilean earthquake. *PLoS ONE*, *11*(7), e0157910. https://doi.org/10.1371/journal.pone.0157910

Williams, D. E., Miller, M. W., Bright, A. J., & Cameron, C. M. (2014). Removal of corallivorous snails as a proactive tool for the conservation of acroporid corals. *PeerJ*, *2*, e680. https://doi.org/10.7717/peerj.680

SUPPORTING INFORMATION

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

How to cite this article: Christie AP, Amano T, Martin PA, Shackelford GE, Simmons BI, Sutherland WJ. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *J Appl Ecol*. 2019;56:2742–2754. https://doi.org/10.1111/1365-2664.13499

13652664, 2019, 12, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13499 by Imperial College London, Wiley Online Library on [02/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License