NEW ZEALAND JOURNAL OF ECOLOGY

REVIEW

Robustness of field studies evaluating biodiversity responses to invasive species management in New Zealand

Robert B. Allen[1*], David M. Forsyth[2], Darryl I. MacKenzie[3] and Duane A. Peltzer[4]

[1]8 Roblyn Place, Lincoln 7608, New Zealand
[2]Vertebrate Pest Research Unit, NSW Department of Primary Industries, 1447 Forest Road, Orange, NSW 2800, Australia
[3]Proteus, Outram 9019, New Zealand
[4]Manaaki Whenua Landcare Research, Lincoln 7608, New Zealand
*Author for correspondence (Email: roballen1706@gmail.com)

Published online: 27 October 2022

Abstract: Benefits of invasive species management for terrestrial biodiversity are widely expected and promoted in New Zealand. Evidence for this is presented in policy and scientific reviews of the literature, but the robustness and repeatability of the underpinning evidence-base remains poorly understood. We evaluated the design of field-based studies assessing biodiversity responses to invasive species management in 155 peer-reviewed articles published in 46 journals from 2010–2019. Each study was assessed against nine principles of experimental design, covering robustness of sampling and avoidance of bias. These principles are important in New Zealand to detect treatment effects from environmental variability driven by underlying gradients such as soil fertility, climate, and disturbance. Across all publications, about half defined the sampling universe (52%) or were unreplicated (54%), whereas most (74%) did not representatively collect data across the sampling universe. Management treatments were specified, with or without only influencing the target species, in 68% of publications. Relatively few studies quantified invasive species (15%) and biodiversity responses (27%) representatively within replicates. Initial conditions and accounting for the effects of experimental implementation were not used in 57% and 84% of publications, respectively. No publications avoided observer/analyst bias using blinding methods, despite this being widely adopted in other scientific fields. We used ordinal logistic regression to understand how these principles varied among categories of biodiversity responses and for major groups of invasive species. Our findings suggest that greater attention to experimental design principles is desirable: supported by researchers, funding agencies, reviewers, and journal editors. Greater resources are not necessarily a solution to these design issues. One alternative is undertaking fewer studies that are individually more expensive because they better adhere to experimental design principles. The challenges of meeting experimental design principles suggests a significant role for other approaches such as systematic monitoring and natural experiments, although many of the design principles we discuss still apply. Our intent in this article is to improve the robustness of future field studies for at least some principles. Robust designs have enduring value, reduce uncertainty, and increase our understanding of when, where and how often the impacts of invasive species on biodiversity are reversible.

Keywords: evidence-base, experimental design, logistic regression, pest control, randomisation, replication, representativeness, sampling universe, systematic review

Introduction

Biodiversity is declining at global, regional, and national scales (Dirzo et al. 2014; Díaz et al. 2019). Non-native invasive species are widely thought to be a major driver of this decline (Vitousek et al. 1997; Vilà et al. 2011; Doherty et al. 2016; but see Gurevitch and Padilla 2004). Beginning with Darwin's visit in 1835, New Zealand has been considered a global exemplar of biodiversity decline caused by invasive species (Thompson 1922; Elton 1958; Allen & Lee 2006; Norton 2009; Simberloff 2019). Māori, as tangata whenua, have long expressed concerns about the plight of biodiversity and the undesirable role of invasive species (Lyver et al. 2008; Harmsworth & Awatere 2013). As a consequence, there is strong societal and political support for management to reverse biodiversity decline. Safeguarding indigenous biodiversity is enshrined in national legislation (e.g. the Conservation Act 1987) and international obligations (e.g. the Convention on Biological Diversity 1993). This has led to some invasive species being managed for benefits to terrestrial biodiversity (Allen & Lee 2006; Jones & McGlinchy 2016; Hulme 2020). In recent efforts, the priority for expenditure has been mammal predator management and impact assessment (Peltzer et al. 2019; Hulme 2020). The Department of Conservation managed

DOI: https://dx.doi.org/10.20417/nzjecol.47.3503

mammal predator and weeds on c. 2.3 and 0.4 million ha, respectively, in the financial year ending June 2019 alone (Department of Conservation 2019). An underlying assumption for management is that biological invasions have caused declines in indigenous biodiversity and that controlling invaders will reverse these effects.

Benefits to biodiversity from invasive species management are widely anticipated in both policy (Parliamentary Commissioner for the Environment 2011; Predator Free 2020; Hackwell & Robinson 2021) and science reviews (Remeš et al. 2012; Byrom et al. 2016; Nelson et al. 2019; Binny et al. 2021; but see Veblen and Stewart 1982; Caughley 1983; Hare et al. 2019). Moreover, these reviews, along with other sources of evidence, are required now for decisions on invasive species management. The robustness of studies underpinning these evidential reviews has received little scrutiny (but see Simpkins et al. 2018), although some previous studies have noted deficiencies (Clayton & Cowan 2010; Smith et al. 2017) and recommended that greater robustness of evidence is desirable (Ferraro & Pattanayak 2006; Reddiex & Forsyth 2006; Doherty & Ritchie 2017; O'Grady 2020). More generally, the reliability of scientific evidence published in peer-reviewed journals requires scrutiny in several disciplines (medicine, Ioannidis 2005; biology, Baker 2016; social science, Camerer et al. 2018). New Zealand is rapidly scaling up invasive species management, as exemplified by Predator Free 2050 (Peltzer et al. 2019) and the National Wilding Conifer Control Programme (Dickie et al. 2022), and hence it is timely to assess the robustness of evidence used to infer biodiversity benefits from invasive species management.

Research evidence can be characterised as: anecdotes and casual observation; logical argument and mathematical modelling; systematic monitoring; and manipulative experiments (McArdle 1996; Hurlbert 1984). Manipulative experiments implemented as, for example, randomised control designs in human medical research have often been encouraged as the "gold standard" for causal inference (Jones & Podolsky 2015). In such experiments, ideally all variables are controlled for (accounted for by non-treatment measurements) and none are uncontrolled (McArdle 1996). If non-treatments work as predicted, it is possible to infer that the results are due to the effect of the treatment variable alone (Oksanen 2001; Ioannidis 2005); in our case the treatment of interest is an invasive species management action. Robust designs are strongest for resolving when and where the impacts of invasive species are, or are not, reversed by managing invasive species (Coomes et al. 2003; Norton 2009; Doherty & Ritchie 2017).

In this paper we consider the robustness of field studies determining biodiversity responses to invasive species management in New Zealand. We first evaluate how key experimental design principles (hereafter principles; Table 1) have been implemented in recent studies testing the response of indigenous, terrestrial biodiversity (hereafter biodiversity) to manipulation of non-native invasive species (hereafter invasive species). Biodiversity responses included compositional, structural or functional characteristics of terrestrial ecosystems. Some principles underpinning robust designs are widely defined, taught and understood, but it is unclear how often these are applied. The manipulations are usually management actions (hereafter treatment). We evaluate peer-reviewed journal publications (hereafter publications), although we acknowledge they are only one source of evidence. Studies published in the last decade (2010–2019) were identified by systematically searching selected journals, review articles, and using an internet search engine. Each publication was assigned one of three ordinal scores with respect to each of nine design principles (Table 1). The frequency of these scores, across all 155 publications, were used to interpret the robustness of recent field studies. We also determined whether the frequency distributions of scores (hereafter score distribution) for each principle were influenced by broad categories of invasive species or biodiversity responses. One expectation was that invasive species categories receiving high scores for principles would be those allocated relatively large expenditure (e.g. mammal predators). An alternative view is that high scores would be found for those organisms that are less challenging to sample in the field (e.g. stationary plants when compared to mobile birds). Finally, we briefly outline some pathways for improving the design of studies investigating biodiversity responses to invasive species management.

Methods

Selection of recent field studies

We selected publications that adopted a field experiment approach; that is, any field studies that derived a statistical inference from measurements of a biodiversity response to a treatment of one or more invasive species. Not all publications dealt with just one such experiment. Where part of a publication was based upon a field experiment, only that part was evaluated. Inferences were sometimes derived from multiple experiments within a single publication (e.g. experiments at multiple locations); in these few instances, all experiments in that publication were evaluated using one set of scores. Although *in situ* biodiversity responses were most commonly reported, we also included experiments in which biodiversity was added (e.g. re-introductions). Types of treatments included any manipulation that attempted to directly affect invasive species at a particular location for a period of time. The majority of treatments were intended to reduce the abundance of invasive species or extirpate them from a defined area. We included publications in which invasive species were added. In some instances, for example fenced sanctuaries, the treatment occurred before biodiversity responses were measured. We included comparisons of islands with and without invasive species because the absence of invasive species most likely reflects actions previously undertaken to restrict the movement of invasive species.

Our attention focussed on a field experiment approach because it forms a strong basis for inductive reasoning (Deaton & Cartwright 2018). In the first instance, we formed a list of 19 peer-reviewed journals where we considered such experimental work would be published (Appendix S1 in Supplementary Materials). We then assessed each article published in those journals during 2010–2019. From these we identified 99 relevant journal articles for further evaluation. These publications were then augmented by an additional 27 (giving a cumulative total of 126) journal articles, published in the same timeframe, found in the reference lists of 20 recent review articles (Appendix S2) identified by the authors. We then used various invasive taxa, biodiversity response and types of treatment, along with New Zealand, as search terms in Google Scholar at 30 April 2020 to select further journal articles. The search terms used were: weeds, fungi, bacteria, *Phytophthora*, invasive species, mammals, and predators for invasive species; native plants, native invertebrates, fungi, bacteria, native bacteria, native fungi, native species, and biodiversity for

Table 1. Experimental design principles used to evaluate each of 155 peer-reviewed journal publications. Also given are the relevance of each of the nine principles to the robustness of experimental inference, score and criteria used to evaluate each principle (higher scores imply more robust inference) and further details on scoring.

Design principles	Relevance to strong inference	Score and criteria used for each principle	Further details on scoring
Principle 1: Define sampling universe	Explicit spatial delineation of a sampling universe gives certainty around the area to which the inference applies.	1 – sampling universe not defined; 2 – defined, but not depicted spatially to allow sampling; 3 – defined and depicted spatially, which can then allow representative sampling	The score was 3 if the experimental inference was not generalised beyond the samples.
Principle 2: Specify treatment(s)	Detailed and time-bound treatment(s) that only impacts the target invasive species is required for robust inference.	1 – treatment action and impact on target species vague; 2 – action defined or impact only on target species substantiated; 3 – action precisely defined and impact specificity known	When treatment(s) have impacts beyond the target invasive species (e.g. fencing), usually known from elsewhere, the score was always < 3.
Principle 3: Use replication	Treatment(s) and non-treatment replication gives robust inference by taking account of variation in the sampling universe.	1 – no replication of treatment(s) or non-treatment; 2 – treatment(s) having 2–4 replicates or non-treatments 2–4 replicates with the other having ≥ 2 replicates; 3 – ≥ 5 treatment(s) and non-treatment replicates	In a nested design, the score was based upon the minimum number of replicates used at any level for treatment(s) or non-treatment(s).
Principle 4: Representatively sample the universe	Random allocation of treatment(s) and non-treatment (s) to replicates representative of the sampling universe give powerful inference.	1 – subjectively located replicates and allocated treatment(s); 2 – randomly allocated treatment(s) or replicates representatively sampling the universe; 3 – randomly allocated treatment(s) to replicates that representatively sample the universe	Where treatment(s) or non-treatment(s) were not all randomly allocated the score was that for subjectively allocated treatment(s). In a nested design, the score was based upon subjective location or allocation if this was done at any level.
Principle 5: Quantify invasive species representatively within replicates	Robust inference from specific quantitative measurements representatively sample all treatment(s) and non-treatment replicates.	1 – target species qualitatively measured without representative sampling; 2 – species quantitatively measured or representatively sampled; 3 – species measured quantitatively and representatively	Where measurements were in part qualitative or in part not representative then the score was as if all measurements were qualitative or not representative.
Principle 6: Quantify biodiversity representatively within replicates	Specific biodiversity responses measured quantitatively using a representative sample within all treatment(s) and non-treatment replicates give robust inference.	1 – biodiversity responses qualitatively measured without representative sampling; 2 – responses quantitatively measured or representatively sampled; 3 – specific responses measured quantitatively and representatively	Where measurements were in part qualitative or in part not representative then the score was as if all measurements were qualitative or not representative.
Principle 7: Determine initial conditions	Representative measurements of target species and biodiversity response(s) made in all replicates before treatment(s)	1 – Initial condition measurements qualitative at best; 2 – some quantitative or representative measurements made of invasive species or biodiversity; 3 – quantitative and representative measurements of invasive species and biodiversity	Where initial conditions were measured for at least some of the invasive species or biodiversity response(s) then the score was as if all initial conditions were measured.
Principle 8: Account for effects of experiment implementation	Robust inference from undertaking procedural controls for both invasive species and indigenous biodiversity.	1 – no procedural controls; 2 – such controls for invasive species treatment(s) and measurements or biodiversity response measurements; 3 – such controls for invasive species and biodiversity	There are many possible procedural controls and the score for this principle is based upon their presence or absence.
Principle 9: Avoid observer/analyst bias	Field measurements and analyses made without bias strengthen inference.	1 – treatments known to all involved; 2 – treatments blind to observers or analysts; 3 – treatments blind to all involved	Score is based upon text noting that blinding was undertaken.

biodiversity responses; and herbicide, biocontrol, control, fungicide, and antibacterial as types of treatment. These were used in various combinations (Appendix S2) and were broad terms to overcome the many possibilities (e.g. number of plant and bird taxa), but also to identify publications of less studied taxa (e.g. bacteria and fungi). From the Google Scholar search we selected a further 7 (total of 133) journal articles. Finally, by searching the reference lists of all 133 journal articles we identified another 22 journal articles to evaluate (for a total of 155 evaluated; Appendix S1). The compiled list thus includes many of New Zealand's field-based publications in journals on invasive species management and biodiversity responses over the past decade.

Scope of selected publications

The 155 publications were found in 46 peer-reviewed journals, with *New Zealand Journal of Ecology* the most common journal (23% of publications; Appendix S1). Mammals as predators were the most commonly-studied invasive species category (58%) and native birds were the most commonly-studied biodiversity response (43%). The most commonly-studied relationships were between mammals as predators and birds (39%), and between mammals as herbivores and plants (19%). The first of these reflects a New Zealand priority of restoring threatened bird species (Hulme 2020). No study reported on potential relationships between mammal herbivores and bird responses despite these relationships being considered important elsewhere (Cocquelet et al. 2019; Crystal-Ornelas et al. 2021), and only two publications reported on relationships between invasive weeds and birds (Table 2).

Choice of principles

Each publication was evaluated against nine experimental design principles selected because of their contribution to the robustness of inference (Table 1; Scheiner & Gurevitch 2001). This list of principles was developed using two approaches. We first surveyed ecological syntheses and reviews that appraised experimental designs but found these only focussed on relatively few principles (Hairston 1989; McArdle 1996; Hone 2007; Dickie et al. 2018). We then surveyed a wider literature to identify additional, relevant principles (Schulz et al. 2002; Ioannidis 2005; Greenlees et al. 2006). The nine emergent principles embody the majority of those pertinent to the design of ecological experiments. Furthermore, considering a greater number of principles than has typically been included in the evaluation of experimental design avoids potential bias from any one or few principles and broadens the range of principles considered in ecological research.

Analyses

To determine the robustness of recent field studies, scores for all nine principles were assigned to each publication (Appendix S1), by one author (RBA), following an initial calibration among authors. We did this to ensure consistency of scoring (Baker 2016). Scores were made using an ordinal scale from 1 to 3 for each design principle, determined using the criteria and details in Table 1. The scores given for a principle were the lowest value justified by the publication. For example, it was assumed that treatments were not randomly allocated unless the article stated otherwise. The score distribution, for each principle, was determined across all 155 publications, where a high score (i.e. 3) implies more robust inference (Tables 1 and 3). Being across all publications means our score distribution for each principle represent the breadth of evidence.

We then assessed whether the score distributions were influenced by invasive species or biodiversity responses as covariates. Each covariate had four categories. Mammal predator, mammal herbivore, and vascular plant weeds (hereafter weed) were three categories used for invasive species, with bird, invertebrate, and vascular plant (hereafter plant) as three categories for biodiversity responses. Invasive species and biodiversity response covariates also included a fourth category, 'other'. Other was a combination of two or three of the previously described categories, not one of the previously described categories, or a combination of previously described categories with one(s) not previously described. The other category for biodiversity responses included non-taxa related characteristics of ecosystems (e.g. soil properties). The score distributions for each of these four categories were based upon ≥ 13 publications and thus also represent a breadth of evidence (Table 2). As this breadth of evidence was not available for rarely studied invasive species or biodiversity responses our results are likely biased against recently established invasive species. An invasive mammal species was included as herbivore mammal if a publication focussed on plant responses but as a predator if about animal responses. These categories defined invasive species and biodiversity response covariates which could explain variation in the score distribution for each principle. When presenting the results from the analyses of these covariates, one category is regarded as a reference category, and effect sizes for the other categories are estimated as the relative difference between that category and the reference category (on an appropriate scale; see below). 'Mammal predator' and 'bird' were regarded as the reference categories for the invasive species and biodiversity response covariates, respectively, because they had the largest numbers of published studies.

Ordinal logistic regression was used to assess the effects of the two covariates on the score distributions for principles 1–6, and regular logistic regression for principles 7–8 (see code in Appendix S3). Ordinal logistic regression can be used to assess the effects of covariates on the three scores, as there is a natural order to scores (Hosmer et al. 2013). The effects of the covariates can be interpreted in terms of odds ratios, in the same manner as for logistic regression. Regular logistic

Table 2. Number of peer-reviewed journal publications classified by invasive species and biodiversity response categories.

Invasive species	**Biodiversity response**				
	Bird	**Invertebrate**	**Plant**	**Other**	**Total**
Mammal predator	60	6	0	24	90
Mammal herbivore	0	0	29	1	30
Weed	2	6	10	3	21
Other	4	1	4	5	14
Total	66	13	43	33	155

regression was used for principles 7 and 8, by combining the frequencies of scores 2 and 3, so that the score distributions had two values (i.e. score 1 and score 2/3 combined). The two scores were combined as there were very few publications with a score of 3 for principles 7 and 8. The data for principle 9 was not analysed further as all publications were assigned a score of 1 (Table 3).

Three models were fit to the data for each of the principles 1–8: no covariate effects; invasive species effects; and, biodiversity response effects. Models with both invasive species and biodiversity response effects were not considered as the covariates were not orthogonal (i.e. 'Mammal predator' invasive species ware primarily studied in relation to 'Bird' and 'Other' biodiversity responses, while 'Mammal herbivore' were almost exclusively studied in relation to 'Plant' biodiversity responses (Table 2). Akaike Information Criterion (AIC) was used as an estimator of prediction error and, given a collection of models, enables the best model(s) to be identified (Burnham & Anderson 2002). Models were ranked using AIC to evaluate the level of support for each model and the relative importance of each covariate to each principle (Table 4). Models with a small difference in AIC (ΔAIC; relative to the top-ranked model) have a similar level of support to the top-ranked model, while models with larger ΔAIC values have much less support. A small difference would be in the range of 0–2 AIC units, and a large difference > 4 AIC units. ΔAIC for the top-ranked model will always equal 0. When using logistic regression methods, estimated effect sizes may be interpreted as odds ratios (i.e. the multiplicative effect on the odds of a publication being given a particular ordinal score). The odds ratio was calculated from the logistic regression coefficient beta where the odds ratio equals exp(beta). If beta equals 0 then the odds ratio equals 1, which is interpreted as no effect. For a covariate category (e.g. weed), an odds ratio not equal to 1 gives the relative number of publications that have a higher score, compared to the lowest score, than that shown by the relative number of publications in the reference category (e.g. mammalian predator for invasive species). If the odds ratio is > 1 then more publications have a higher score and if the odds ratio is < 1 then less publications have a higher score. The odds ratios and overlap in associated 95% confidence intervals (CI) generated by the two forms of logistic regression were used to compare the effect of invasive species or biodiversity response categories on score distributions for each principle.

Robustness of recent field studies

One model had more support than the other models considered for each principle, except for principles 4 and 7 (Table 4). For these principles, two models had similar levels of support. The model with the biodiversity response covariate most supported was for principle 1, and the model with the invasive species covariate most supported was for principles 2, 3, 5,

Table 3. Percentage of 155 peer-reviewed journal publications receiving each score for nine experimental design principles. A higher score implies more robust inference.

Principle	Score		
	1	2	3
Principle 1: Define sampling universe	48	25	27
Principle 2: Specify treatment(s)	32	56	12
Principle 3: Use replication	54	28	18
Principle 4: Representatively sample the universe	74	17	9
Principle 5: Quantify invasive species representatively within replicates	50	35	15
Principle 6: Quantify biodiversity representatively within replicates	5	68	27
Principle 7: Determine initial conditions	57	38	5
Principle 8: Account for effects of experiment implementation	84	15	1
Principle 9: Avoid observer/analyst bias	100	0	0

Table 4. Relative difference in Akaike Information Criterion value (ΔAIC) for each model compared to the highest-ranked model for the eight principles analysed by the two forms of logistic regression.

Principle	Model		
	No covariates	Invasive species effects only	Biodiversity response effects only
Principle 1: Define sampling universe	3.94	2.42	0.00
Principle 2: Specify treatment(s)	5.19	0.00	6.29
Principle 3: Use replication	44.70	0.00	17.69
Principle 4: Representatively sample the universe	0.00	1.05	4.42
Principle 5: Quantify invasive species representatively within replicates	9.74	0.00	9.33
Principle 6: Quantify biodiversity representatively within replicates	31.33	0.00	11.46
Principle 7: Determine initial conditions	0.33	0.00	2.50
Principle 8: Account for effects of experiment implementation	7.58	0.00	2.95

6, and 8. The model with no covariate effects was highest ranked for principle 4, although the model with the invasive species covariate also had some support. For principle 7, the invasive species effects model was ranked highest, but the model with no covariates had a very similar level of support. We now consider the most supported model within a wider consideration of the rationale for and evaluation of results for each principle.

Define sampling universe

Principle 1 provides objectivity for the area to which the inference applies (Table 1; McArdle 1996). This universe can be expressed spatially at a point in time (e.g. locality, species distribution, or island) and, if the boundary is carefully mapped with known accuracy, gives robust spatial limits. About half (48%) of publications had the lowest score because they did not define a sampling universe (Table 3). The plant category scored lowest of the biodiversity responses, and the bird category tended to have relatively high scores (Fig. 1). The high score for bird possibly reflects publications in which the birds were studied on easily defined areas such as islands (e.g. Jones et al. 2015) or an area fenced to exclude predators (e.g. Bogisch et al. 2016). The utility of publications which do not define the sampling universe is restricted, because their results cannot be interpreted in relation to an area over which they apply or extrapolated to larger areas (Smith et al. 2017). Such a limitation appears common in ecological research. Dickie et al. (2018), for example, found that 92% of 75 DNA-based biodiversity studies from around the world did not define a sampling universe; often studies described sampling locations in detail, but not how these locations were chosen to be representative of any larger area.

Specify treatment(s)

A time-bounded treatment should be defined with sufficient methodological detail to provide clarity about what is being tested; moreover, the treatment should only impact the target invasive species and not be confounded with other actions (principle 2). Fifty-six percent of publications received an intermediate score for this principle, usually because non-target invasive species or biodiversity itself were potentially influenced by the treatment(s) (Bellingham et al. 2016; van Vianen et al. 2018), rather than the treatment action not being defined (Tables 1 and 3). The use of non-target specific poisons as treatment(s) for mammal predators and mammal herbivores may explain why their categories have lower scores than weed or other invasive species categories (Fig. 2). Moreover, field experiments were at times confounded by the addition of biodiversity to some invasive species treatments or replicates: these included additions of native species of birds (Bombaci et al. 2018), plants (Graham et al. 2013) and reptiles (Tanentzap and Lloyd 2017). In such cases, any biodiversity response, in certain treatment(s) or replicates, may be directly, or indirectly, a consequence of the addition(s). Some experiments were also confounded by the treatment(s) being applied previously, at least in part, to non-treatment(s) (Fea & Hartley 2018), or the treatment(s) changing during the course of a study. Invasive mammal species treatment(s), for example, sometimes change trapping and poisoning regimes during a study, potentially altering both effectiveness of control and non-target impacts (Parkes & Murphy 2003). There is the potential for hidden treatments to create complex interactions that drive ecosystem-level responses (Wardle et al. 2012). As a consequence, scores for this principle were high when a manipulation added only the target invasive species (Pawson et al. 2010) or affected only the target species. Morgan et al. (2012) reported the timing, intensity, and constraints when using magpie (*Gymnorhina tibicen*)-specific traps, and so received the highest possible score for this principle.

Use replication

Treatment and non-treatment replicates are required to adequately account for variation within a sampling universe (principle 3; Hurlbert 1984; Chalcraft 2019). Replication is challenging for some field experiments and the level of replication required for strong inference can be clarified through power analyses (Carpenter 1989; Eberhardt & Thomas 1991; Allen et al. 2003). Such analyses were rarely undertaken. Fifty-four percent of publications evaluated had no replication of treatment(s) or non-treatment(s) (Table 3). Twelve percent of publications considered one treatment location, such as an island or fenced sanctuary, without including non-treatment locations. This may be why mammal predator publications, often treated at such locations, had relatively low scores (Fig. 2). While it is particularly challenging to replicate mammal herbivore studies they were replicated as well as weed studies (Fig. 2). Relatively few replicates could be required if there is little variation in treatment(s) and non-treatment(s) (Oksanen

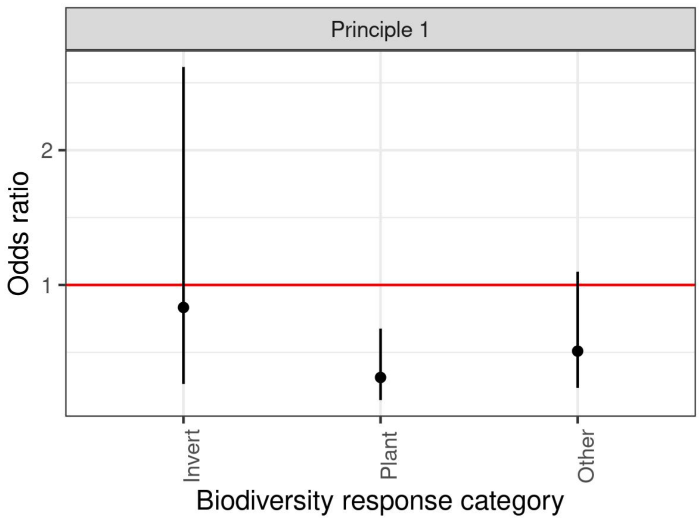

Figure 1. Estimated odds ratios (and 95% confidence intervals) for biodiversity response categories. The best supported model for Principle 1: Define sampling universe was for the biodiversity response covariate (Tables 1, 2 and 4). The reference category was bird (horizontal line). If the odds ratio for a category was > 1 then more publications of that category have a higher score (effect) than the reference category (bird), and if the odds ratio was < 1 then fewer publications of that category have a higher score than the reference category. Invert = invertebrates.

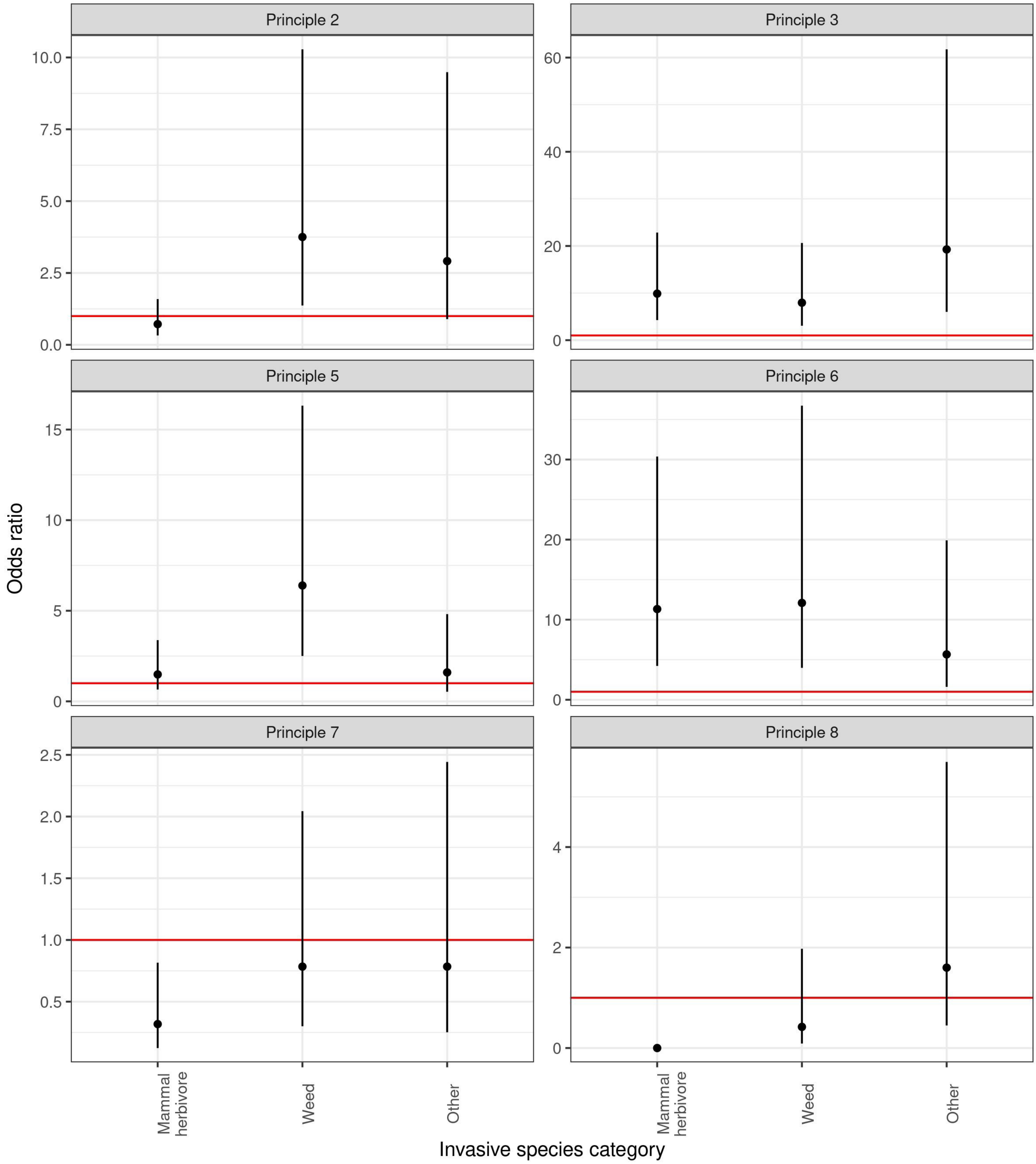

Figure 2. Estimated odds ratios (and 95% confidence intervals) for invasive species categories. The best supported models for Principles 2: Specify treatment(s), 3: Use replication, 5: Quantify invasive species representatively within replicates, 6: Quantify biodiversity responses representatively within replicates, 7: Determine initial conditions, and 8: Account for the effects of experiment implementation was the invasive species covariate (Tables 1, 2 and 4). The reference category was mammal predator (horizontal line). If the odds ratio for a category was > 1 then more publications of that category have a higher score (effect) than the reference category (mammal predator), and if the odds ratio was < 1 then fewer publications of that category have a higher score than the reference category.

2001). However, New Zealand is remarkably diverse in climate, soils, geology, and disturbance history, even at small spatial scales (Wardle 1991). For example, fresh leaf phosphorus (P) concentrations subjectively sampled in Te Urewera, central North Island, captured > 90% of the global variation in leaf concentrations (Richardson et al. 2008). This variation reflected topographically related differences in soils at the scale of hundreds of meters. Not only does this small-scale variation drive marked differences in nutrient cycling (Richardson et al. 2004), but also habitat use by invasive species such as feral pigs (*Sus scrofa*) (Forsyth et al. 2016). Elsewhere, Forsyth et al. (2015) have suggested, through a combination of empirical studies and modelling, that the long-term impacts of invasive deer and rodents on indigenous forests will be greater on sites with relatively high soil P availability than nearby sites with low P availability. Hence, robust replication and randomisation are needed (Hurlbert 1984; Filazzola & Cahill 2021). Chance intrusions also occurred in some replicates during an experiment, which added variability to the data. Such intrusions can include flood damage (Simpkins et al. 2015), predation by domestic dogs (*Canis familiaris*) (Robertson et al. 2019), build-ups in disease causing fungi (Perrott & Armstrong 2011), and variation in avian malaria (*Plasmodium* spp.) infection rates (Alley et al. 2010).

Representatively sample the universe

It is desirable that both treatment and non-treatment replicates, which can occur at multiple levels in a complex design, are representative of the sampling universe. This is usually achieved by random or systematic location of experimental replicates, with treatment(s) randomly allocated among those replicates (principle 4; Mentges et al. 2021). Seventy-four percent of publications did not have replicates that representatively sampled a sampling universe, nor employed a random allocation of treatment(s), whereas 9% of publications did both (Fig. 3; Table 3). It is also desirable to have treatment(s) and non-treatment(s) replicates intermixed within the sampling universe and independent, i.e. sufficiently distant, to avoid spatial autocorrelation or spill-over of treatment effects (Hurlbert 1984), which appears to be rarely achieved in practice. Some publications acknowledged interchange between treatment and non-treatment replicates (Parlato & Armstrong 2013).

Management imperatives or financial and logistical constraints have sometimes driven suboptimal treatment(s) allocation (Parkes et al. 2006; Broadbent et al. 2017). Subjectivity accentuates potential problems of hidden treatment effects and sampling bias (Smith et al. 2017). If this bias is systematic, then they become embedded in meta-analyses and reviews. Peltzer et al. (2014) shed some light on an apparent systematic bias in fenced exclosure studies that explore biodiversity responses to removing invasive ungulate herbivores (Fig. 3). Conservation managers subjectively established c. 100 fenced exclosures throughout New Zealand's indigenous forests during the 1970s and 1980s to remove

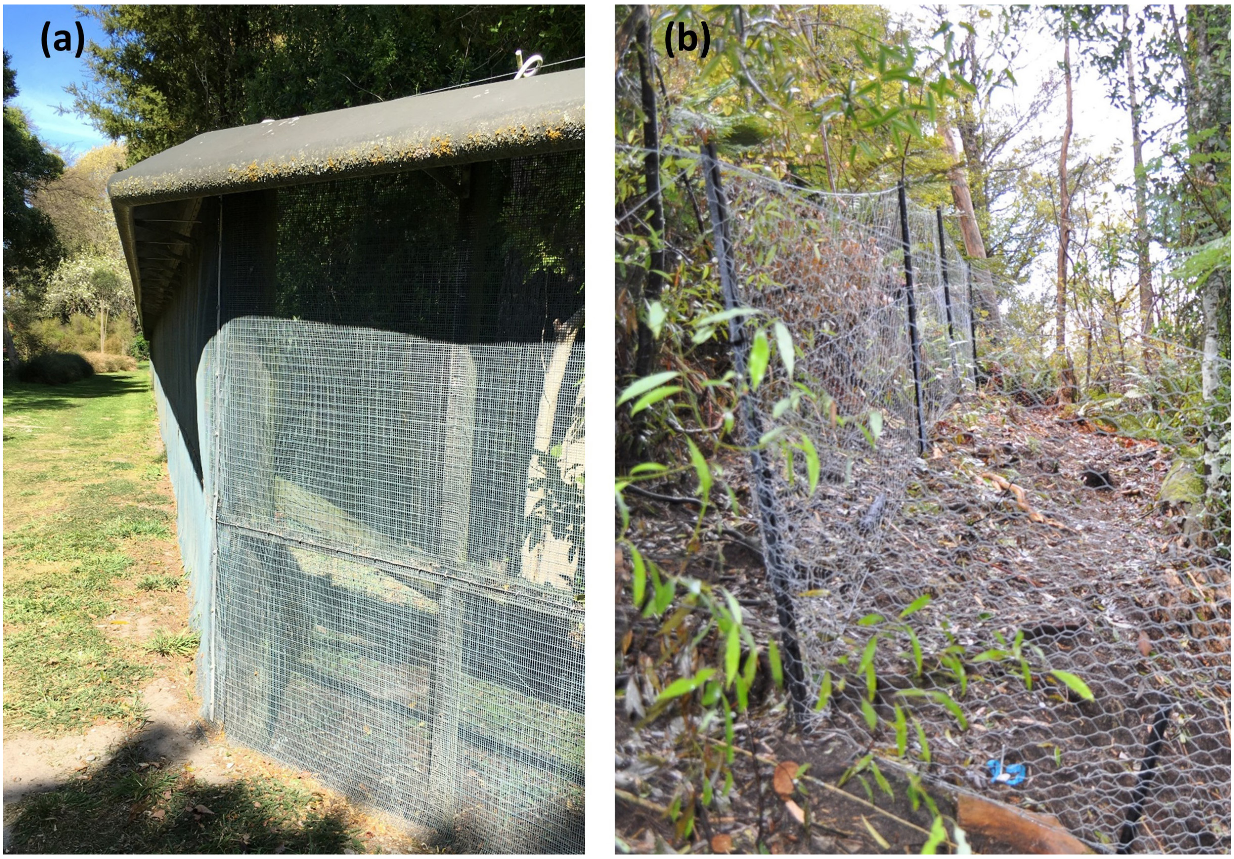

Figure 3. A fence constructed with a fine wire mesh, capped, and topped with an electric wire has been used at Riccarton Bush, Christchurch, to exclude invasive mammal predators such as brushtail possums, stoats and rats (a). Simpler fences constructed with a coarse wire mesh have been used in Te Urewera to exclude invasive mammal herbivores such as deer (b). Such fencing treatments are widely deployed for animal exclusion but often exclude non-target species, have been located subjectively, treat a small area, are subject to periodic breaches, and potentially influence ecosystem-level responses (e.g. seed production and litter fall).

ungulate browsing in the forest understorey (Mason et al. 2010). As expected, palatable tree species showed higher numbers of small trees inside exclosures than outside (Mason et al. 2010). Unexpectedly, the numbers of these palatable species found at a representative grid of locations (unfenced) nationally were similar to those found inside the fenced exclosures (Peltzer et al. 2014). It appears that the exclosures were constructed, knowingly or unknowingly, on sites where ungulate impacts were strongest. Under such circumstances, any level of replication (number of fenced exclosures), for example, would be overridden by sampling bias.

Quantify invasive species representatively within replicates

Representative sampling of replicates and quantitative measures of invasive species (principle 5) ensures robust inferences (Oksanen 2001; Salo et al. 2010; Webster & Rutz 2020). Jones et al. (2013), for example, received the highest score for this principle because skink (*Oligosoma* spp.) abundance was determined from a grid of pit-fall traps that sampled each replicate. In treatments where invasive species have previously been eradicated (e.g. islands), our evaluation of principle 5 partly used support from existing literature cited, rather than publication-specific measurements. Fifty percent of publications reviewed neither representatively sampled nor quantitatively measured the target invasive species, whereas 15% of publications did both (Table 3). We had expected that weeds would score relatively high for principle 5 (Fig. 2), because such stationary organisms are relatively easy to measure. In some publications, quantitative measurement was not undertaken because the target invasive species was assumed to be eradicated or never present; this assumption needs to be verified, because ongoing (re)invasion occurs and can be difficult to detect (Salo et al. 2010; Bellingham et al. 2010; Drummond & Armstrong 2019). In other publications, the manipulation itself was assumed to cause an invasive species difference between treatment(s) and non-treatment(s) replicates. However, Forsyth et al. (2013) showed that there can be large uncertainty around the impacts of an invasive species manipulation in treatment and non-treatment replicates. Salo et al.'s (2010) review stressed the importance of measuring the abundance of predators before and during an experiment to be sure about their impact on prey.

Quantify biodiversity responses representatively within replicates

Representative sampling of replicates and quantitative measures of biodiversity responses (principle 6) also ensures that inferences are robust (Oksanen 2001; Salo et al. 2010; Webster & Rutz 2020). Sixty-eight percent of publications received an intermediate score for this principle (Table 3). This was because publications were often based upon quantitative measurements of biodiversity responses but not representative sampling. In quantifying biodiversity responses, mammal predator studies scored relatively low when compared with the scores for the other three invasive species categories (Fig. 2). Twenty-seven percent of publications described representative sampling (e.g. plots) and quantitative measures within replicates (Table 3). Representative sampling remains poorly used, even in the face of recent technological developments that simplify its use (e.g. Geographic Information Systems; Smith et al. 2017). Rather, subjective sampling was the norm in our reviewed publications, and this has important implications. Avoiding certain locations or features within replicates may reduce variability, and likely increases the probability of significant effects (Dickie et al. 2018). For example, McAlpine et al. (2016) subjectively avoided sampling within invasive pine (*Pinus contorta*) slash sites in treatments when measuring native seedling establishment. Such an approach means that results are representative of only a subset of the sampling universe, and this reduced sampling universe should be defined.

Determine initial conditions

If few replicates are used in an experiment, or background variation among replicates is large, then inferences can be strengthened by accounting for initial conditions. One way of doing this is by using before-and-after treatment(s) measurements of invasive species or biodiversity responses (principle 7; Hurlbert 1984; Hairston 1989; Salo et al. 2010). In our evaluation, before treatment measurements were interpreted as representing initial conditions rather than non-treatments. Innes et al. (2012) received the highest score for this principle because they undertook pre-treatment magpie counts in all treatment and non-treatment replicates using a grid of stations. In 57% of publications, initial conditions were determined qualitatively at best and were not measured representatively (Table 3). Mammal herbivore publications received a relatively low number of high scores (Fig. 2). Measuring initial conditions is particularly challenging and expensive for mammal herbivores with large home ranges. Where initial conditions were measured, sometimes seasonal measurements were more restricted before a treatment than after (van Vianen et al. 2018), or for only a subset of biodiversity responses (Robertson et al. 2019). Representative and quantitative measurements of initial conditions were made in 5% of studies for both biodiversity responses and invasive species (Table 3). Indeed, manipulations of invasive species were at times made long before any measurements, with long-term vegetation change creating a challenge for understanding any variability in initial conditions (Holdaway et al. 2014).

Account for effects of experiment implementation

Procedural controls ensure that the observed effect is not an artefact of some aspect(s) of experiment implementation (Underwood 1997). These were evaluated for both target invasive species and biodiversity responses (principle 8). Procedural controls were rarely used (84% received the lowest score; Table 3). Mammal herbivore publications scored relatively low for principle 8 when compared to the other three invasive species categories (Table 3; Fig. 2). Non-target impacts of various poisons, used to manipulate invasive mammals, on biodiversity has forced some scrutiny of the unintended poisoning of native animals as a procedural control (Kemp et al. 2019). Another example examined whether bird translocations to islands without predators caused a directional selection on a stress response to capture (Adams et al. 2013). Publications with procedural controls tended to publish those results alone, without including the invasive species and biodiversity responses to treatment(s). Only the study by Wardle et al. (2010) included more than one of the many possible procedural controls. The importance of procedural controls has long been recognised in ecological research (Hurlbert 1984; Greenlees et al. 2006) yet is rarely used compared with its use in medical research (Price et al. 2008). There are many possible procedural controls and individual studies need to consider what effects to account for in experimental implementation.

Avoid observer/analyst bias

We include treatments being blind to observers and analysts (principle 9; Schulz et al. 2002) as our final principle. Implementing and reporting experiments that are blind is a major approach used to avoid bias (Ioannidis 2005), and such biases may be particularly important when an agency is assessing the effectiveness of it's own management. All publications used field experiments in which the treatment(s) were apparently known to the observer and analyst (Table 3); this could introduce bias in data collection, analysis or interpretation, particularly when, for example, treatment effects on biodiversity are not quantified but subjectively estimated. The absence of observer/analyst blinding contrasts with its common use in other research areas as a means of overcoming bias (Schulz et al. 2002). Aside from using blinding approaches, independent audits of field measurements could be used to help overcome observer bias.

Pathways for improvement

Enhancing field studies

Experiments require substantive assumptions, prior information, and are not independent of "expert" knowledge (Deaton & Cartwright 2018). We support a greater commitment to robust experimental designs but recognise that logistical and resource constraints often determine the design (Christie et al. 2019; Filazzola & Cahill 2021). As a consequence, our intent is to improve the robustness of future field studies for at least some principles. Crawley (2015) offers some guidance and considered randomisation and replication as the two essential concepts in experimental design. We discuss below examples of how some principles could be better implemented in field studies.

Clearer protocols for some principles

Some design principles such as replication and randomisation are well established and could readily be applied more widely. For example, defining a sampling universe can be done with little expense and tests of invasive species management could more often use random allocation of treatments to overcome systematic bias. Other principles such as procedural controls and minimising observer/analyst bias are more challenging to utilise in ecological research. Implementing these challenging principles requires a deconstruction of the experiment to understand what is required. Implementing an experimental manipulation could create, for example, artefacts that directly or indirectly affect the target invasive species, other invasive species, or biodiversity responses. Not all of these will necessarily be important for procedural controls, and could change through time. Procedural controls, and specifying the treatment, can be improved when the target invasive species is not removed but added as a treatment. Testing whether implementing a manipulation itself directly influences an invasive species may require non-treatment(s) to receive only the action used in a treatment, for example, shooting at an invasive ungulate species using blanks in the non-treatment areas in an experiment testing the effects of aerial shooting. Procedural controls are challenging in a common New Zealand manipulation, fencing, whereby there is a direct effect on the target invasive species as well as other invasive species. Fences were commonly used to exclude invasive mammal predators to restore bird populations, but they often also exclude invasive herbivores (Fig. 3). If the herbivore exclusion enabled forest understories to become denser, then reduced visibility may mean less bird nest predation and increased bird abundance (Cocquelet et al. 2019; Crystal-Ornelas et al. 2021). While biodiversity managers do need to operate at times without knowledge of underlying causal processes, we suggest that understanding such relationships provides insights into both ecological processes and the effectiveness of both current or new management interventions on biodiversity.

Greater scrutiny pre-implementation

Some mechanisms already exist for greater scrutiny before a field experiment is implemented. Pre-registering a study is now becoming an accepted part of conservation science (Parker et al. 2019; Filazzola & Cahill 2021). Pre-registration is the archiving of a detailed description of a proposed study's research questions/hypotheses, experimental design, and data collection and analysis methods in a public registry (Nosek et al. 2015). Registered reports are similar to pre-registration but are subject to peer review at a journal prior to the study beginning. If after peer review the journal editor is satisfied with the logic and design of the proposed study, then in-principal acceptance of the study is given, regardless of results, provided that the robust design and analysis described in the registered report were followed. Some ecological journals offer the registered report option (e.g. *Conservation Biology*). While such options may not yet be appropriate for all studies, it is likely that pre-registration and registered reports will become an increasingly accepted way of doing ecological research and are one mechanism to help ensure that key design principles are addressed in field studies.

Adequately resourced experiments

A lack of resources in ecological research and monitoring can lead to a weak and biased evidence base (Dirzo et al. 2014; Christie et al. 2019). This limits our ability to provide robust recommendations to decision-makers. Greater resources is not the only solution, and one alternative is to undertake fewer experiments that are individually more expensive because they better adhere to design principles (Baker 2016; Christie et al. 2019; Filazzola & Cahill 2021). For example, this may allow replication (principle 3) and quantification of biodiversity responses (principle 6) to be strengthened in mammal predator studies as well as initial conditions (principle 7) and procedural controls (principle 8) to be more commonly applied in mammal herbivore studies (Fig. 2). Fewer experiments could simply be generated by, for example, using registered reports, as defined above, to improve and filter designs prior to treatment or data collection.

Strengthen publication processes

Reviewers, editors and journals could be more insistent that publications meet widely known and accepted design principles (Filazzola & Cahill 2021). Our evaluation suggests that design robustness may not be a key decisive factor when manuscripts are accepted for publication in a wide range of journals. We suggest that citations of individual publications be qualified by denoting their critical design principle limitations (e.g. subjectively sampled or unreplicated). When meta-analyses are undertaken, weighting systems can be used to give greater influence to individual studies having more robust designs (Stewart 2010; Christie et al. 2019). However, increasingly sophisticated analyses cannot remove potentially widespread systematic bias (Peltzer et al. 2014). Increased transparency about robustness, uncertainties or bias could strengthen their inclusion in decision-making and, moreover, reinforce the utility of having robust experiments as an evidence base.

Adopting other approaches

The challenges of meeting design principles when implementing robust experiments, as an inductive approach, suggests a significant role for other approaches (Eberhardt and Thomas 1991; Oksanen 2001; Lyver et al. 2008; Davies & Gray 2015; Jones & Podolsky 2015; Deaton & Cartwright 2018; Kreyling et al. 2018; Munafò & Smith 2018; Filazzola & Cahill 2021; Ockendon et al. 2021). This includes roles for big-data mining, the complementary role of deductive reasoning, traditional knowledge, systematic monitoring, and natural experiments. When management imperatives determine that studies are undertaken at large spatial scales it can be difficult to replicate an experimental manipulation and this reinforces a case for using systematic monitoring (Hurlbert 1984; Eberhardt & Thomas 1991). Natural experiments might occur only at one location but still provide opportunities for new evidence that would be impossible to obtain otherwise, particularly when initial conditions are known (principle 7; Ockendon et al. 2021; see also Diamond 1983; Davies & Gray 2015). We emphasise that most of the design principles outlined above (e.g. sampling universe, representativeness, and observer/analyst bias) also apply when generating evidence from these other approaches. It is for such reasons that the Department of Conservation's Biodiversity Monitoring and Reporting System samples New Zealand's indigenous forests nationally on a grid with a random starting point (Allen et al. 2003; Forsyth et al. 2018; Bellingham et al. 2020). Such approaches can generate insights at larger spatial and temporal scales than experimental studies, thereby increasing the generalisability of results, but also have a limited ability to disentangle different drivers of biodiversity responses. It is therefore desirable to include interpretive covariates in systematic monitoring studies (Allen et al. 2003), although experiments do not necessarily relieve us of this need (e.g. Deaton & Cartwright 2018). Interpretive covariates focus our attention away from typical effects shown in experimental studies onto research that accommodates idiosyncrasies and differences (Kreyling et al. 2018). Context matters for invasive species impacts (Sapsford et al. 2020). With such approaches, alternative analytical methods such as noise clustering of community composition (Wiser & de Cáceres 2013) and statistical matching (e.g. propensity scoring) can be used, for example, to investigate the effects of manipulating invasive species on biodiversity (Ramsey et al. 2019). These methods require some treatment(s) locations to have covariate values that overlap with non-treatment locations. This overlap is achieved for common invasive species and biodiversity when sampled at large spatial scales, as well as capturing spatial structure (Legendre et al. 2004).

Concluding comment

Our review reveals that both long-recognised (e.g. replication and randomisation; Hurlbert 1984) and more recent (e.g. blinding; Schulz et al. 2002) principles of design have often not been adopted in ecological field experiments on invasive species and biodiversity. The consequences of this for the reliability of the evidence-base remains to be investigated, but we suggest that a prudent response is to improve, where possible, the robustness of studies. Other methodological areas not considered here, such as statistical analyses choices, incorporating measurement error, role of audit, utility of response variables, code sharing and data accessibility, as well an open publication process (Wiser et al. 2001; Salo et al. 2010; Fanelli 2012; Tressoldi et al 2013; Nosek et al. 2015; Fraser et al. 2018; Mason et al. 2018; Cusser et al. 2021; Filazzola & Cahill 2021; Etherington et al. 2022; Wagenmakers et al. 2022), also underpin the utility and robustness of publications. We emphasise here that robustness of design, and thus quality of evidence, precedes or underpins the methodological areas mentioned above and ultimately the effectiveness of management interventions for altering biodiversity. The importance of this issue cannot be overstated: declines in many components of biodiversity continue despite more than a century of management efforts. Increasingly robust studies will allow us to better understand when, where, and how often the impacts of invasive species will persist or be influenced by management (Veblen & Stewart 1982; Coomes et al. 2003; Peltzer et al. 2019; Simberloff 2019; Hulme 2020).

Acknowledgements

We thank Chris Jones and the Predator Free New Zealand Trust for providing access to an unpublished report. This article was funded by the authors and Department of Conservation. John Herbert, Sharyn Laing, Murray Llewellyn, and Elaine Wright provided motivation for this review and Sarah Richardson the photograph of a fence excluding invasive ungulate herbivores. Comments by James Russell and several anonymous reviewers greatly improved this manuscript.

Data and code availability

All data are provided in Appendix S1 and the code in Appendix S3.

Author contributions

All authors were involved in conceptualisation, RBA and DMF developed the methods and DIM the analysis. RBA wrote the original manuscript and DMF, DIM, and DAP undertook reviews and editing. RBA and DAP completed the visualisation.

References

Adams NJ, Parker KA, Cockrem JF, Brunton DH, Candy EJ 2013. Inter-island differences in the corticosterone responses of North Island Saddlebacks (*Philesturnus rufusater*) in New Zealand do not suggest selective effects of translocation. Emu-Austral Ornithology 113: 45–51.

Allen RB, Lee WG eds. 2006. Biological invasions in New Zealand. Heidelberg, Springer. 457 p.

Allen RB, Bellingham PJ, Wiser SK 2003. Developing a forest biodiversity monitoring approach for New Zealand. New Zealand Journal of Ecology 27: 207–220.

Alley MR, Hale KA, Cash W, Ha HJ, Howe L 2010. Concurrent avian malaria and avipox virus infection in translocated South Island saddlebacks (*Philesturnus carunculatus carunculatus*). New Zealand Veterinary Journal 58: 218–223.

Baker M 2016. 1,500 scientists lift the lid on reproducibility. Nature 533(7604): 452.

Bellingham PJ, Towns DR, Cameron EK, Davis JJ, Wardle DA, Wilmshurst JM, Mulder CP 2010. New Zealand island restoration: seabirds, predators, and the importance of history. New Zealand Journal of Ecology 34: 115–136.

Bellingham PJ, Kardol P, Bonner KI, Buxton RP, Morse CW, Wardle DA 2016. Browsing by an invasive herbivore promotes development of plant and soil communities during primary succession. Journal of Ecology 104: 1505–1517.

Bellingham PJ, Richardson SJ, Gormley AM, Allen RB, Cook A, Crisp PN, Forsyth DM, McGlone MS, McKay M, MacLeod CJ, van Dam-Bates P, Wright EF 2020. Implementing integrated measurements of Essential Biodiversity Variables at a national scale. Ecological Solutions and Evidence 1: e12025.

Binny RN, Innes J, Fitzgerald N, Pech R, James A, Price R, Gillies C, Byrom AE 2021. Long-term biodiversity trajectories for pest-managed ecological restorations: eradication vs. suppression. Ecological Monographs 91: p.e01439.

Bogisch M, Cree A, Monks JM 2016. Short-term success of a translocation of Otago skinks (*Oligosoma otagense*) to Orokonui Ecosanctuary. New Zealand Journal of Zoology 43: 211–220.

Bombaci S, Pejchar L, Innes J 2018. Fenced sanctuaries deliver conservation benefits for most common and threatened native island birds in New Zealand. Ecosphere 9: p.e02497.

Broadbent AA, Orwin KH, Peltzer DA, Dickie IA, Mason NW, Ostle NJ, Stevens CJ 2017. Invasive N-fixer impacts on litter decomposition driven by changes to soil properties not litter quality. Ecosystems 20: 1151–1163.

Burnham KP, Anderson DR 2002. Model selection and inference: a practical information-theoretic approach. 2nd edn. New York, Springer-Verlag. 488 p.

Byrom AE, Innes J, Binny RN 2016. A review of biodiversity outcomes from possum-focused pest control in New Zealand. Wildlife Research 43: 228–253.

Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T, Altmejd A 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2: 37–644.

Carpenter SR 1989. Replication and treatment strength in whole-lake experiments. Ecology 70: 453–463.

Caughley G 1983. The deer wars: the story of deer in New Zealand. Auckland, Heinemann. 187 p.

Chalcraft R 2019. To replicate, or not to replicate—that should not be a question. Ecology Letters 22: 1174–1175.

Christie AP, Amano T, Martin PA, Shackelford GE, Simmons BI, Sutherland WJ 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56: 2742–2754.

Clayton R, Cowan P 2010. Management of animal and plant pests in New Zealand—patterns of control and monitoring by regional agencies. Wildlife Research 37: 360–371.

Cocquelet A, Mårell A, Bonthoux S, Baltzinger C, Archaux F 2019. Direct and indirect effects of ungulates on forest birds' nesting failure? An experimental test with artificial nests. Forest Ecology and Management 437: 148–155.

Coomes DA, Allen RB, Forsyth DM, Lee WG 2003. Factors preventing the recovery of New Zealand forests following control of invasive deer. Conservation Biology 17: 450–459.

Crawley MJ 2015. Statistics: an introduction using R. 2nd ed. New Jersey, USA, John Wiley & Sons Ltd. 339 p.

Crystal-Ornelas R, Brown JA, Valentin RE, Beardsley C, Lockwood JL 2021. Meta-analysis shows that overabundant deer (Cervidae) populations consistently decrease average population abundance and species richness of forest birds. Ornithological Applications 123: duab040.

Cusser S, Helms IV J, Bahlai CA, Haddad NM 2021. How long do population level field experiments need to be? Utilising data from the 40-year-old LTER network. Ecology Letters 24: 1103–1111.

Davies GM, Gray A 2015. Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). Ecology and Evolution 5: 5295–5304.

Deaton A, Cartwright N 2018. Understanding and misunderstanding randomized controlled trials. Social Science and Medicine 210: 2–21.

Department of Conservation 2019. Department of Conservation Te Papa Ātawhai annual report for the year ended June 2019. Wellington, Department of Conservation.

Diamond JM 1983. Ecology: Laboratory, field and natural experiments. Nature 304: 586–587.

Díaz S, Settele J, Brondízio ES, Ngo HT, Agard J, Arneth A, Balvanera P, Brauman KA, Butchart SH, Chan KM, Garibaldi LA, Liu J, Subramanian SM, Midgley GF, Miloslavich P, Molnár Z, Obura D, Pfaff A, Polasky S, Purvis A, Razzaque J, Reyers B, Chowdhury RR, Shin Y-J, Visseren-Hamakers I, Willis KJ, Zayas CN 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366(6471): eaax3100.

Dickie IA, Boyer S, Buckley HL, Duncan RP, Gardner PP, Hogg ID, Holdaway RJ, Lear G, Makiola A, Morales SE, Powell JR 2018. Towards robust and repeatable sampling methods in eDNA-based studies. Molecular Ecology Resources 18: 940–952.

Dickie IA, Sprague R, Peltzer DA, Green J, Orwin K, Sapsford S 2022. Beyond control: Applying ecological research to improve long-term outcomes of wilding conifer management. New Zealand Journal of Ecology 47: 3475.

Dirzo R, Young HS, Galetti M, Ceballos G, Isaac NJB, Collen B 2014. Defaunation in the Anthropocene. Science 345: 401–406.

Doherty TS, Ritchie EG 2017. Stop jumping the gun: a call for evidence-based invasive predator management. Conservation Letters 10: 15–22.

Doherty TS, Glen AS, Nimmo DG, Ritchie EG, Dickman CR 2016. Invasive predators and global biodiversity loss. Proceedings of the National Academy of Sciences of the United States of America 113: 11261–11265.

Drummond FM, Armstrong DP 2019. Use of distance sampling to measure long-term changes in bird densities in a fenced wildlife sanctuary. New Zealand Journal of Ecology 43: 3371.

Eberhardt LL, Thomas JM 1991. Designing environmental field studies. Ecological Monographs 61: 53–73.

Elton CS 1958. The ecology of invasions by animals and plants. London, Methuen. 181 p.

Etherington TR, Brock JMR, Perry GLW, Wyse SV 2022. Implementing the Transparency and Openness Promotion Guidelines for data and code to support computational

reproducibility within the New Zealand Journal of Ecology. New Zealand Journal of Ecology 46: 3460.
Fanelli D 2012. Negative results are disappearing from most disciplines and countries. Scientometrics 90: 891–904.
Fea N, Hartley S 2018. The balancing act of nest survival: survival of a small endemic bird in the face of ship rat predation and other risk factors. Avian Conservation and Ecology 13(2): 11.
Filazzola A, Cahill Jr JF 2021. Replication in field ecology: Identifying challenges and proposing solutions. Methods in Ecology and Evolution 12: 1780–1792.
Ferraro PJ, Pattanayak SK 2006. Money for nothing? A call for empirical evaluation of biodiversity conservation investments. PLoS Biol 4: e105.
Forsyth DM, Ramsey DS, Veltman CJ, Allen RB, Allen WJ, Barker RJ, Jacobson CL, Nicol SJ, Richardson SJ, Todd CR 2013. When deer must die: large uncertainty surrounds changes in deer abundance achieved by helicopter-and ground-based hunting in New Zealand forests. Wildlife Research 40: 447–458.
Forsyth DM, Wilson DJ, Easdale T, Kunstler G, Canham CD, Ruscoe WA, Wright EF, Murphy L, Gormley A, Gaxiola A, Coomes DA 2015. Century-scale effects of invasive deer and rodents on the dynamics of forests growing on soils of contrasting fertility. Ecological Monographs 85: 157–180.
Forsyth DM, Allen RB, Allen RKJ, Affeld K, MacKenzie DI 2016. Soil phosphorus predicts feral pig (*Sus scrofa*) occupancy, detection probability and feeding activity in a temperate montane rainforest. Wildlife Research 43: 277–287.
Forsyth DM, Ramsey DS, Perry M, McKay M, Wright EF 2018. Control history, longitude and multiple abiotic and biotic variables predict the abundances of invasive brushtail possums in New Zealand forests. Biological Invasions 20: 2209–2225.
Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F 2018. Questionable research practices in ecology and evolution. PloS one 13(7): e0200303.
Graham M, Veitch D, Aguilar G, Galbraith M 2013. Monitoring terrestrial bird populations on Tiritiri Matangi Island, Hauraki Gulf, New Zealand, 1987–2010. New Zealand Journal of Ecology 37: 359–369.
Greenlees MJ, Brown GP, Webb JK, Phillips BL, Shine R 2006. Effects of an invasive anuran [the cane toad (*Bufo marinus*)] on the invertebrate fauna of a tropical Australian floodplain. Animal Conservation 9: 431–438.
Gurevitch J, Padilla DK 2004. Are invasive species a major cause of extinctions? Trends in Ecology and Evolution 19: 470–474.
Hackwell K, Robinson M 2021. Protecting our natural ecosystems' carbon sinks. Wellington, Forest and Bird. 49 p.
Hairston NG 1989. Ecological experiments: purpose, design and execution. New York, Cambridge University Press. 384 p.
Hare KM, Borrelle SB, Buckley HL, Collier KJ, Constantine R, Perrott JK, Watts CH, Towns DR 2019. Intractable: species in New Zealand that continue to decline despite conservation efforts. Journal of the Royal Society of New Zealand 49: 301–319.
Harmsworth GR, Awatere S 2013. Indigenous Māori knowledge and perspectives of ecosystems. In: Dymond J ed. Ecosystem services in New Zealand: conditions and trends. Palmerston North, Manaaki Whenua Press. Pp. 274–286.
Holdaway RJ, Rose AB, Newell CL, Carswell FE 2014. Demographic drivers of biomass carbon recovery in degraded perennial tussock grassland, with and without domestic grazing. New Zealand Journal of Ecology 38: 201–212.
Hosmer DW, Lemeshow S, Sturdivant RX 2013. Applied logistic regression. New Jersey, John Wiley and Sons. 510 p.
Hone J 2007. Wildlife damage control. Canberra, CSIRO. 171 p.
Hulme PE 2020. Plant invasions in New Zealand: global lessons in prevention, eradication and control. Biological Invasions 22: 1539–1562.
Hurlbert SH 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54: 187–211.
Innes JG, Spurr EB, Arnold G, Morgan DK, Waas JR, Watts C 2012. Using five-minute bird counts to study magpie (*Gymnorhina tibicen*) impacts on other birds in New Zealand. New Zealand Journal of Ecology 36: 324–332.
Ioannidis JPA 2005. Why most published research findings are false. PLoS Medicine 2(8): e124.
Jones CJ, McGlinchy A 2016. Costs of control of vertebrate predators by New Zealand regional and national agencies: July 2010–June 2015. Landcare Research Contract Report LC2658. Lincoln, Landcare Research. 23 p.
Jones C, Norbury G, Bell T 2013. Impacts of introduced European hedgehogs on endemic skinks and weta in tussock grassland. Wildlife Research 40: 36–44.
Jones CJ, Lyver P, Whitehead A, Forrester G, Parkes J, Sheehan M 2015. Grey-faced petrel (*Pterodroma gouldi*) productivity unaffected by kiore (Pacific rats, *Rattus exulans*) on a New Zealand offshore island. New Zealand Journal of Zoology 42: 131–144.
Jones DS, Podolsky SH 2015. The history and fate of the gold standard. The Lancet 385(9977): 1502–1503.
Kemp JR, Mosen CC, Elliott GP, Hunter CM, van Klink P 2019. Kea survival during aerial poisoning for rat and possum control. New Zealand Journal of Ecology 43: 3351.
Kreyling J, Schweiger AH, Bahn M, Ineson P, Migliavacca M, Morel-Journel T, Christiansen JR, Schtickzelle N, Larsen KS 2018. To replicate, or not to replicate–that is the question: how to tackle nonlinear responses in ecological experiments. Ecology Letters 21: 1629–1638.
Legendre P, Dale MR, Fortin MJ, Casgrain P, Gurevitch J 2004. Effects of spatial structures on the results of field experiments. Ecology 85: 3202–3214.
Lyver POB, Taputu TM, Kutia ST, Tahi B 2008. Tūhoe Tuawhenua mātauranga of kererū (*Hemiphaga novaseelandiae novaseelandiae*) in Te Urewera. New Zealand Journal of Ecology 32: 7–17.
Mason NW, Peltzer DA, Richardson SJ, Bellingham PJ, Allen RB 2010. Stand development moderates effects of ungulate exclusion on foliar traits in the forests of New Zealand. Journal of Ecology 98: 1422–1433.
Mason NW, Holdaway RJ, Richardson SJ 2018. Incorporating measurement error in testing for changes in biodiversity. Methods in Ecology and Evolution 9: 1296–1307.
McAlpine KG, Howell CJ, Wotton DM 2016. Effects of tree control method, seed addition, and introduced mammal exclusion on seedling establishment in an invasive *Pinus contorta* forest. New Zealand Journal of Ecology 40:

302–309.
McArdle BH 1996. Levels of evidence in studies of competition, predation and disease. New Zealand Journal of Ecology 20: 7–15.
Mentges A, Blowes SA, Hodapp D, Hillebrand H, Chase JM 2021. Effects of site-selection bias on estimates of biodiversity change. Conservation Biology 35: 688–698.
Morgan DK, Waas JR, Innes JG, Arnold G 2012. Native bird abundance after Australian magpie (*Gymnorhina tibicen*) removal from localised areas of high resource availability. New Zealand Journal of Ecology 36: 333–339.
Munafò MR, Smith GD 2018. Repeating experiments is not enough. Nature 553(401): 399–401.
Nelson NJ, Briskie JV, Constantine R, Monks J, Wallis GP, Watts C, Wotton DM 2019. The winners: species that have benefited from 30 years of conservation action. Journal of the Royal Society of New Zealand 49: 281–300.
Norton DA 2009. Species invasions and the limits to restoration: learning from the New Zealand experience. Science 325(5940): 569–571.
Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M 2015. Promoting an open research culture. Science 348: 1422–1425.
Ockendon N, Amano T, Cadotte M, Downey H, Hancock MH, Thornton A, Tinsley-Marshall P, Sutherland WJ 2021. Effectively integrating experiments into conservation practice. Ecological Solutions and Evidence 2: e12069.
O'Grady C 2020. Ecologists push for more reliable research. Science 370(6522): 1260–1261.
Oksanen L 2001. Logic of experiments in ecology: is pseudoreplication a pseudoissue? Oikos 94: 27–38.
Parker T, Fraser H, Nakagawa S 2019. Making conservation science more reliable with preregistration and registered reports. Conservation Biology 33: 747–750.
Parkes JP, Murphy E 2003. Management of introduced mammals in New Zealand. New Zealand Journal of Zoology 30: 335–359.
Parkes JP, Robley A, Forsyth DM, Choquenot D 2006. In my opinion: adaptive management experiments in vertebrate pest control in New Zealand and Australia. Wildlife Society Bulletin 34: 229–236.
Parlato EH, Armstrong DP 2013. Predicting post-release establishment using data from multiple reintroductions. Biological Conservation 160: 97–104.
Parliamentary Commissioner for the Environment 2011. Evaluating the use of 1080: predators, poisons and silent forests. Wellington, Parliamentary Commissioner for the Environment.
Pawson SM, McCarthy JK, Ledgard NJ, Didham RK 2010. Density-dependent impacts of exotic conifer invasion on grassland invertebrate assemblages. Journal of Applied Ecology 47: 1053–1062.
Peltzer DA, Allen RB, Bellingham PJ, Richardson SJ, Wright EF, Knightbridge PI, Mason NWH 2014. Disentangling drivers of tree population size distributions. Forest Ecology and Management 331: 165–179.
Peltzer DA, Bellingham PJ, Dickie IA, Houliston G, Hulme PE, Lyver POB, McGlone M, Richardson SJ, Wood J 2019. Scale and complexity implications of making New Zealand predator-free by 2050. Journal of the Royal Society of New Zealand 49: 412–439.
Perrott JK, Armstrong DP 2011. *Aspergillus fumigatus* densities in relation to forest succession and edge effects: implications for wildlife health in modified environments. EcoHealth 8: 290–300.
Predator Free 2020. Predator Free 2050 Vision. https://predatorfreenz.org/big-picture/pf-2050-vision/ (accessed 4 May 2020).
Price DD, Finniss DG, Benedetti F 2008. A comprehensive review of the placebo effect: recent advances and current thought. Annual Review of Psychology 59: 565–590.
Ramsey DSL, Forsyth DM, Wright E, McKay M, Westbrooke I 2019. Using propensity scores for causal inference in ecology: options, key considerations and a case study. Methods in Ecology and Evolution 10: 320–331.
Reddiex B, Forsyth DM 2006. Control of pest mammals for biodiversity protection in Australia. II. Reliability of knowledge. Wildlife Research 33: 711–717.
Remeš V, Matysioková B, Cockburn A 2012. Nest predation in New Zealand songbirds: exotic predators, introduced prey and long-term changes in predation risk. Biological Conservation 148: 54–60.
Richardson SJ, Peltzer DA, Allen RB, McGlone MS, Parfitt RL 2004. Rapid development of phosphorus limitation in temperate rainforest along the Franz Josef soil chronosequence. Oecologia 139: 267–276.
Richardson SJ, Allen RB, Doherty JE 2008. Shifts in leaf N:P ratio during resorption reflect soil P in temperate rainforest. Functional Ecology 22: 738–745.
Robertson HA, Guillotel J, Lawson T, Sutton N 2019. Landscape-scale applications of 1080 pesticide benefit brown kiwi (*Apteryx mantelli*) and fantails (*Rhipidura fuliginosa*) at Tongariro Forest, New Zealand. Notornis 66: 1–15.
Salo P, Banks PB, Dickman CR, Korpimäki E 2010. Predator manipulation experiments: impacts on populations of terrestrial vertebrate prey. Ecological Monographs 80: 531–546.
Sapsford SJ, Brandt AJ, Davis KT, Peralta G, Dickie IA, Gibson RD, Green JL, Hulme PE, Nuñez MA, Orwin KH, Pauchard A 2020. Towards a framework for understanding the context dependence of impacts of non-native tree species. Functional Ecology 34: 944–955.
Scheiner SM, Gurevitch J eds. 2001. Design and analysis of ecological experiments. 2nd edn. Oxford, Oxford University Press. 445 p.
Schulz KF, Chalmers I, Altman DG 2002. The landscape and lexicon of blinding in randomized trials. Annals of Internal Medicine 136: 254–259.
Simberloff D 2019. New Zealand as a leader in conservation practice and invasion management. Journal of the Royal Society of New Zealand 49: 259–280.
Simpkins C, Perry GL, Glaser A, Allerby T, Dennis TE 2015. Effects of predation by introduced mammals and mortality due to severe floods on population viability of the endangered Blue Duck (*Hymenolaimus malacorhynchos*). Emu-Austral Ornithology 115: 146–157.
Simpkins CE, Lee F, Powers BF, Anderson S, Asena Q, Brock JMR, Perry GLW 2018. Population viability analyses in New Zealand: a review. New Zealand Journal of Ecology 42: 107–115.
Smith ANH, Anderson MJ, Pawley MDM 2017. Could ecologists be more random? Straightforward alternatives to haphazard spatial sampling. Ecography 40: 1251–1255.
Stewart G 2010. Meta-analysis in applied ecology. Biology Letters 6: 78–81.

Tanentzap AJ, Lloyd KM 2017. Fencing in nature? Predator exclusion restores habitat for native fauna and leads biodiversity to spill over into the wider landscape. Biological Conservation 214: 119–126.
Thompson GM 1922. The naturalisation of animals and plants in New Zealand. Cambridge, Cambridge University Press. 606 p.
Tressoldi PE, Giofré D, Sella F, Cumming G 2013. High impact = high statistical standards? Not necessarily so. PLOS One 8: e56180.
Underwood AJ 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge, Cambridge University Press. 504 p.
van Vianen J, Burge OR, MacFarlane AT, Kelly D 2018. The effects of single aerial 1080 possum-control operations on common forest birds in the South Island, New Zealand. New Zealand Journal of Ecology 42: 169–178.
Veblen TT, Stewart GH 1982. The effects of introduced wild animals on New Zealand forests. Annals of the Association of American Geographers 72: 372–397.
Vilà M, Espinar JL, Hejda M, Hulme PE, Jarošík V, Maron JL, Pergl J, Schaffner U, Sun Y, Pyšek P 2011. Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. Ecology Letters 14: 702–708.
Vitousek PM, D'Antonia CM, Loope LL, Rejmanek M, Westbrooks RG 1997. Introduced species: a significant component of human-induced global change. New Zealand Journal of Ecology 21: 1–16.
Wagenmakers E-J, Sarafoglou A, Aczel B 2022. One statistical analysis must not rule them all. Nature 605: 423–425.
Wardle DA, Karl BJ, Beggs JR, Yeates GW, Williamson WM, Bonner KI 2010. Determining the impact of scale insect honeydew, and invasive wasps and rodents, on the decomposer subsystem in a New Zealand beech forest. Biological invasions 12: 2619–2638.
Wardle DA, Bellingham PJ, Fukami T, Bonner KI 2012. Soil-mediated indirect impacts of an invasive predator on plant growth. Biology Letters 8: 574–577.
Wardle P 1991. Vegetation of New Zealand. Cambridge, Cambridge University Press. 672 p.
Webster MM, Rutz C 2020. How STRANGE are your study animals? Nature 582: 337–340.
Wiser SK, Bellingham PJ, Burrows LE 2001. Managing biodiversity information: development of New Zealand's National Vegetation Survey databank. New Zealand Journal of Ecology: 25: 1–17.
Wiser SK, De Cáceres M 2013. Updating vegetation classifications: an example with New Zealand's woody vegetation. Journal of Vegetation Science 24: 80–93.

Received: 11 March 2022; accepted: 1 June 2022
Editorial board member: George Perry

Supplementary material

Additional supporting information may be found in the supplementary material file for this article:

Appendix S1. Score for each of nine design principles by publication (n = 155) and the year of publication. One of four invasive species categories and one of four biodiversity response categories are also given for each publication.

Appendix S2. Twenty recent review articles are given where the reference lists were searched for relevant publications. Also given are the search terms used in Google Scholar to select further publications.

Appendix S3. Code for logistic regression analyses completed in this study.

The New Zealand Journal of Ecology provides supporting information supplied by the authors where this may assist readers. Such materials are peer-reviewed and copy-edited but any issues relating to this information (other than missing files) should be addressed to the authors.