Received: 4 January 2024 | Revised: 21 February 2024 | Accepted: 25 February 2024

DOI: 10.1111/ele.14400

ECOLOGY LETTERS cnrs WILEY

EDITORIAL

Principles of experimental design for ecology and evolution

Dustin J. Marshall

School of Biological Sciences, Monash University, Melbourne, Victoria, Australia

Correspondence
Dustin J. Marshall, School of Biological Sciences, Monash University, Melbourne, VIC, Australia.
Email: dustin.marshall@monash.edu

Editor: Jonathan M. Chase

Abstract

Good experimental design is critical for sound empirical ecology and evolution. However, many contemporary studies fail to replicate at the appropriate biological or organizational level, so causal inference might have less vigorous support than often assumed. Here, I provide a guide for how to identify the appropriate scale of replication for a range of common experimental designs in ecological and evolutionary studies. I discuss the merits of replicating multiple scales of biological organization. I suggest that experimental design be discussed in terms of the scale of replication relative to the scale at which inferences are sought when designing, discussing and reviewing experiments in ecology and evolution. I also suggest that more conversations about experimental design are needed, and I hope this piece stimulates such conversation.

KEYWORDS
causal inference, experimental design, replication

INTRODUCTION

Good design is an essential component of empirical biology. Without good experimental design, few of the causal inferences that we seek to make about biological phenomena have rigorous support. Experimental design features briefly in most biostatistics courses, but in my experience, researchers at all career stages struggle with a key feature of experimental design: replication.

I have been a senior editor at different journals for more than a decade, and one of the few consistencies I've seen is experimental designs with low inferential power. In fact, I reject a significant proportion (>20%) of manuscripts because they suffer from inadequate experimental design. I use 'inadequate' not as a pejorative term, but more precisely to mean that the experimental design does not support the results or inferences. Often, the design problem occurs because of a lack of replication at the appropriate scale (more on this later). I think more conversations about experimental design are needed, both within and among research teams—my goal here is to prompt such conversations.

I will outline what I believe are the essentials of experimental design, and provide a selection of applications of these principles to specific examples in ecology and evolution. I will focus on examples that include my experiences with inadequate experimental designs. Please note that none of these examples represents a particular case that I have handled as an editor; rather, I have chosen hypothetical examples that capture the essence of what has appeared many times from many different researchers.

When covering the essentials of experimental design, I will avoid discussing specific statistical approaches, even though I believe that good experimental design and clear analytical plans go hand-in-hand. Statistical best practices are changing so rapidly that I risk focusing on soon-to-be obsolete methods, whereas good experimental design is eternal. A good experimental design must have at least two attributes: (1) Replication at the appropriate biological scale; (2) Factors of interest be free from confounding so their influence can be disentangled from each other and other factors. These attributes seem easy to obtain, but they can be more elusive than many people realize. I will expand on each of these principles in turn before exploring specific examples.

REPLICATION AT THE APPROPRIATE SCALE

Whenever we ask a biological question, we are asking whether a pattern is meaningful, or the difference is large

relative to some baseline level of variation. We need to replicate in order to discern whether our factor of interest explains more variation than the variation that occurs independently of that factor of interest. The problem is that there are many scales of organization that we could use to get our baseline understanding of variation—this is where things get difficult.

Baseline estimates of variation must come from replication at the same scale as the factor of interest. Other scales of replication can occur and even be desirable, but the critical level of replication is that exact scale at which we apply our factor of interest. To illustrate using a very simple example, imagine studying the insect communities on the leaves of a plant and how nutrient additions affect these communities. Individual plants receive either nutrients or a control substance—the scale at which the factor of interest is applied is 'individual plant', and so we must replicate at that scale. Taking several leaves from a single treatment plant and several leaves from a single control plant and treating those leaves as replicates is clearly inappropriate. In this case, we would be using a baseline level of variation that exists within plants to examine whether the difference caused by the treatment between plants is substantial—a clear mismatch in the scales of variation that we are comparing. Likewise, imagine we were interested in the effect of a pollutant on the metabolic rate of a fish, and we exposed individual fish to either the pollutant or a control and then measured their metabolic rate. Clearly, using repeated measures of the metabolic rate of only a single pollutant-exposed fish and only a single control fish multiple times is inappropriate—variation over time does not provide estimates of baseline variation that exist between individuals, which is our scale of comparison. These are very obvious examples that most will find trivial. However, it is important to establish some key principles upon which we can all agree so that we can apply these same principles in more complicated scenarios.

Scales of biology and experiments

It is often difficult, unfeasible or inefficient to apply treatments at exactly the same scales as the level at which we are seeking to make inferences. In our plant example, it might be too costly to house each plant individually, or it might be impossible to prevent the nutrient we add to the soil from leaching over to an adjacent plant. Even though we are interested in individual plants, we are compelled to apply the treatment (nutrients in our example) at the scale of plots that contain multiple plants. Because we are applying our factor of interest at the scale of plots, we need a baseline that estimates the variation between plots—plants are not replicates. Even if we measure our response variables on individual plants within each plot, we would then take the mean (or, better yet, use statistical nesting) of each plot as our true level of replication. Likewise, we might apply a water-soluble pollutant to a tank containing five snails per tank; the treatment is applied to tanks, so the level of replication is tank. The individual snails are useful subsamples but not replicates (Figure 1). More generally, mismatches between the biological scale of interest and the scale at which the treatment is applied can accidentally generate inappropriate experimental designs. If your interest is in individuals, but you apply the treatment at a higher scale, replication at that scale should occur.

It is not uncommon for researchers to muddle the scale of replication when their interest is at higher scales of organization, such as populations or species, rather than individuals. Imagine a study investigating whether ocean acidification alters the shells of snails—are individuals that grow in acidic conditions more robust than individuals that grow in normal pH? (Figure 2). As biologists, we recognize that long-term exposures to an environmental stressor might yield evolutionary responses in populations that cannot be predicted from short-term exposures. So, we might use naturally occurring CO_2 seeps or upwelling regions as our driver of low pH; we are using differences among populations to make our inferences. Our treatment (lower pH) is being applied at the population level; we therefore need multiple low pH populations and multiple normal pH populations—otherwise, we will be comparing our effect of interest using an inappropriate level of variation (among-individual variation would be used to compare among-population-level differences; Figure 2). Similar issues have been raised in the context of comparative analyses.

Confusing the appropriate scale of replication can lead to the misallocation of effort whereby researchers laboriously measure many subsamples at lower biological scales but only have a single replicate at the appropriate scale—in other words, no replication whatsoever. In such cases, researchers are making inferences based on what is essentially the outcome of a coin flip: a fifty–fifty chance. Because no two populations will be perfectly identical, one must differ from the other (however slightly)—that difference may or may not be driven by our factor of interest, but we cannot know because we lack the appropriate comparator. That difference could well be statistically significant if variance between individuals is used to evaluate whether the difference between populations is relatively small—but the analysis does not answer the question of interest regarding causal effects.

This is an important nuance that must be flagged here. Doing an experiment with two populations of individuals and comparing those populations statistically is perfectly valid—from this, we can infer that the two populations are different. But inferring that a specific difference in the characteristics of the populations drives that difference is invalid—we need to replicate populations that vary in the focal characteristic to make this inference. I have sometimes heard from rejected authors that,

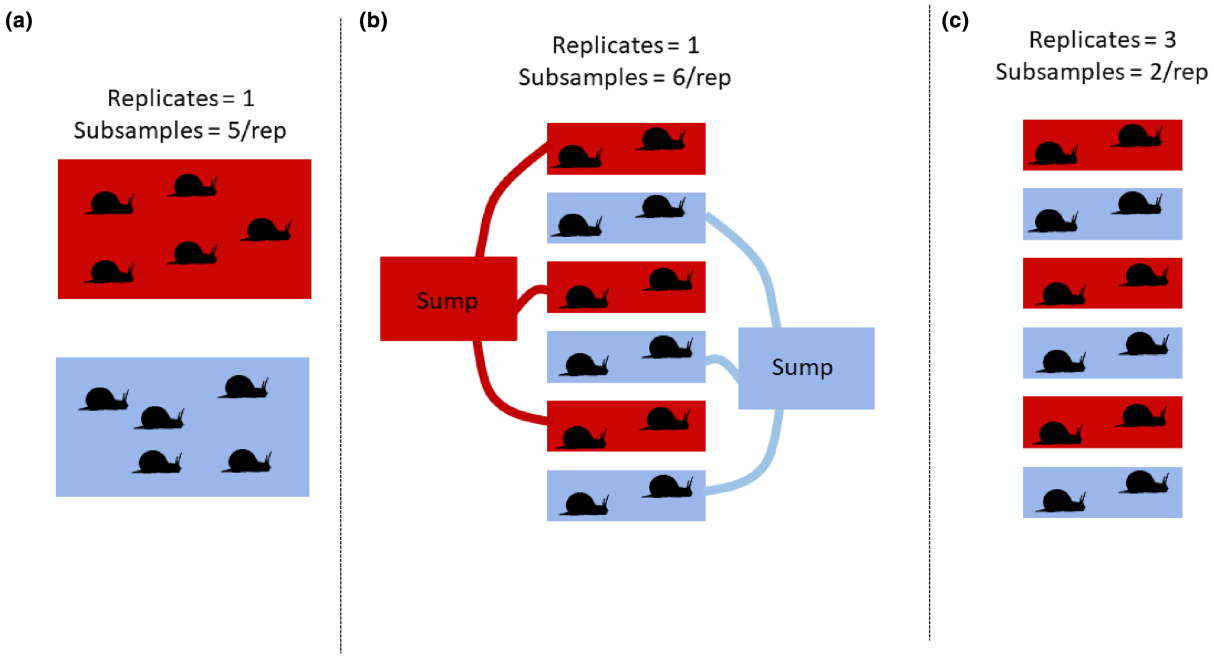


FIGURE 1 Three scenarios for exploring the effect of temperature (represented by blue and red) on a snail phenotype. Scenario (a) shows one tank per temperature, where all five snails are subsamples and are hence unreplicated at the scale at which the treatment is applied. Scenario (b) shows three tanks of each temperature, but each of the tanks is connected to a sump such that a common water supply confounds temperature, and the design is therefore unreplicated. Scenario (c) shows an acceptable design that includes three replicate tanks with two subsamples per replicate.

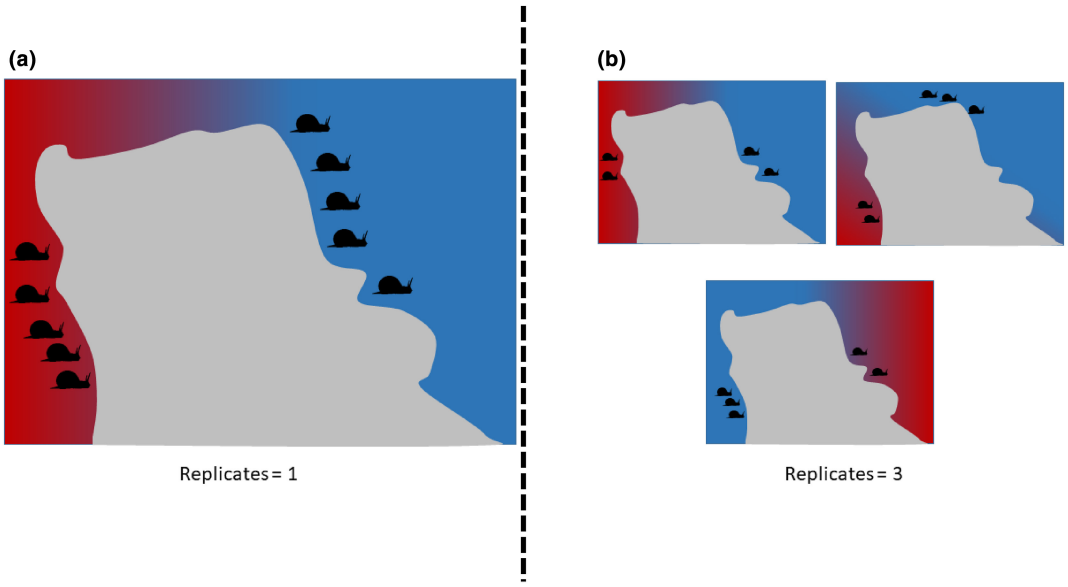


FIGURE 2 Two scenarios for exploring population-level differences in snail phenotype. The colour gradient represents some environmental condition that varies systematically from left to right across a headland. Scenario (a) involves sampling multiple snails from two populations. This approach is fine for inferring whether snails on either side of the headland are different but not why they are different (i.e. we cannot make inferences as to whether the environmental gradient drives any differences). Scenario (b) shows a design with replicate headlands such that we can make inferences about the driver of the differences in the headland. Ideally, the environmental gradient would show different orientations among different headlands so as to rule out confounding directional (i.e. east–west) effects. Note that were we to replace 'population' for 'species' the same problem would arise. Comparing just two species is fine if one is just interested in whether those two species are different, but if we want to determine why they are different (and we usually do), we need to replicate species that vary in the putative driving factor.

because the differences that they detected matched prior expectations, the lack of replication at the appropriate scale is irrelevant. In reality, whether the difference between the populations matches or defies expectations is meaningless because no true estimate of the baseline variation at the appropriate scale has been estimated.

Using grammar to identify scales of replication

How can we avoid misidentifying the scale at which to replicate? A good rule of thumb is to use the grammar of one's description of the study as the clue for how to replicate. For any experiment, we usually use an adjective to describe the factor of interest (temperature, latitude and predation) that precedes a noun (population, species, genotype and individual). In these instances, the adjective identifies the treatment, and the noun identifies the scale at which the treatment was applied, and hence the scale that should be replicated. For example, if you create three warm enclosures and three cool enclosures for an experiment on the effects of temperature on herbivory, then 'warm' and 'cool' are the adjectives, and 'enclosure' is the scale at which the treatment is applied and, therefore, the scale that must be replicated.

The special problem of constant temperature chambers

A swathe of studies manipulate temperature using constant-temperature rooms (or cabinets) but treat individuals within those chambers as replicates. Often, these studies are not replicated at the scale at which the treatment is applied (cabinet) and, therefore, are not generating the appropriate baseline variance for comparison. It is tempting to argue that temperature effects are strong, so we can ignore temperature chamber effects. But the strength of chamber effects relative to temperature effects for any one experiment is unknown. I suspect there are considerable phenotypic differences in organisms grown in different chambers at the same temperature, but this needs testing.

Even if we accept that chamber effects are smaller than temperature effects, our goal is to isolate and estimate temperature effects as free of noise as possible—at best, chamber effects add noise to the data that diminish our effect sizes. While we might have strong (and often correct) expectations about the direction of temperature effects, we remain completely ignorant of the effects of chambers. Each chamber will have its own idiosyncrasies, so we cannot assign an 'average' strength or direction of an effect of chamber in order to correct for these effects. By confounding temperature and chamber, we generate less precise estimates of temperature effects.

It is important to recognize that temperature chambers are expensive and resources are limited—so how do we proceed? Thankfully, there are several options. Assuming researchers have only two chambers, repeating the experiment multiple times and swapping the temperatures between experimental runs will disentangle the effect of chamber from temperature.

A second solution to having a limited number of temperature chambers is to manipulate temperatures within the chamber. In this case, multiple temperatures are applied within the one chamber, thereby 'breaking' the confounding between the treatment of interest and the scale at which it is applied. Practically, it is easier/cheaper to heat than cool, so setting the chamber at the coolest temperature of interest and then creating warmer conditions within subchambers (which are now the unit of replication) by applying a warming device to achieve multiple temperature treatments within the one chamber. In order to avoid confounding, every enclosure should include a warming device, but the subchambers in the ambient (cooler) treatment should leave the warming device switched off.

A third solution to having a limited number of temperature chambers is to treat temperature as a continuous variable. Sometimes called a 'gradient' approach, if we have at least three chambers, we can treat temperature as a continuous variable and examine linear relationships between temperature and the response variable of interest. Note that this approach comes with important caveats: (i) When analysed correctly, this approach is statistically identical to any other regression where n = the number of chambers; (ii) I would refrain from fitting any model other than a linear model in this analysis unless there are many more than three chambers.

Why I prefer not to use the term 'Pseudoreplicates'

The issues I have discussed so far touch on long-standing discussions about what and how to replicate in biology. These discussions are contentious (Hurlbert, 1984; Schank & Koehnle, 2009), and I do not wish to revisit these controversies. However, I do need to mention terminology here—Hurlbert (1984) argued that we should identify true replicates and pseudoreplicates as a way of distinguishing scales at which things are measured—pseudoreplication occurs when there are repeated measures that are not statistically independent. In the framework I described above, pseudoreplicates would be those that occur below the biological scale of organization at which the factor of interest is applied. For example, in Figure 1a, the snails inside the tank would be regarded as pseudoreplicates, whereas the tanks in Figure 1c would be the 'true' replicates. My concern is that, under a pseudoreplication framework, the snails in the tanks in Figure 1c would also be referred to as pseudoreplicates because they are not independent from each other. Instead, I would prefer to regard these snails as useful for increasing the precision of our estimates. More generally,

I prefer to call samples below the level of the factor of interest 'subsamples' rather than pseudoreplicates, primarily because the phrase pseudoreplicate implies that such samples have no value. Instead, I think there are clear instances where replicating at multiple scales is valuable and when subsampling is desirable (I explore these below).

I also personally dislike the intellectual framework that is used to identify pseudoreplicates—which tends to emphasize the 'independence' of replicates. Of course, replicates should be unaffected by each other as much as possible, but it can be difficult to determine objectively where independence begins and ends (e.g. how do we really know that a snail in one tank has absolutely no effect on the phenotype of a snail in another tank?). Oftentimes, we must use our intuition as biologists or best guesses as to how organisms might perceive the world and be affected by it, and what are likely mechanisms by which independence may or may not be maintained. Trying to categorize independence when most things lie on a continuum of dependence is fraught. Others have discussed these issues, and I direct the reader to these papers (Heffner et al., 1996; Schank & Koehnle, 2009). Instead, the goal should be gaining an estimate of variation that is representative of what occurs in the absence of the factor of interest at the appropriate scale.

WHEN TO REPLICATE AT MULTIPLE SCALES

Multilevel designs

In some instances, it is necessary or more efficient to explore factors of interest at multiple biological scales of organization. For example, consider a study on the role of soil type and herbivory on secondary metabolite content in a grass. Rather than artificially manipulate soil type (which could be difficult or unrealistic), one could make use of the natural variation in soil types among different sites. 'Site' is the unit of replication for examining the effects of soil type, but excluding herbivores from entire sites is impractical. Instead, we would benefit from constructing smaller cages (and cage controls) around individual plants within each site—thus, 'plant' is the unit of replication for the herbivory treatment. As long as there are replicates of the lower-scale treatment (herbivory exclusion) across replicate sites of each soil type, multilevel (sometimes called 'partly nested' or 'split-plot') analyses can easily accommodate such designs. Hence, the rule of thumb about identifying scales of replication using the grammar of the experimental description still holds—we have sites with different soil types and plants with different herbivory treatments, so both sites and plants are replicates of each respective treatment.

Transgenerational experiments where parents are exposed to one environment or another while the offspring of those parents are exposed to all possible environments are another increasingly common example of multilevel designs. In this case, parents are the unit of replication for the parental treatment and offspring are often the unit of replication for the offspring treatment (Figure 3). The offspring replicates also have the additional role of acting as subsamples that increase the precision of the estimates of the parental treatment. Note that Figure 3a illustrates an all-too-familiar problem: that designs that completely lack replication can still be a lot of work.

Multilevel designs can have the benefit of allowing us to maximize efficiency—replicating only at the scales we must—and avoid costlier or impractical designs (in the example above, excluding herbivores at the site level). However, it is worth noting that interactions between factors at different levels are not synonymous with interactions between factors applied at the same level, and should not be interpreted as such. For example, an interaction between soil effects at the level of the site and herbivory exclusion at the level of the plant is not the same as an interaction that might occur if we exclude herbivores at the scale of whole sites. Given that much of biology is scale-dependent, we should be cautious about extrapolating from observations of interactions at one scale to interactions at others.

Biological realism or necessity

Sometimes the biology of a system requires that multiple units below the scale at which the factor is applied be included in a single replicate. For example, imagine you are interested in how different phytoplankton species affect the timing of reproduction in copepods. For this experiment, copepods would be placed in containers with phytoplankton species A or B, and so the container is the unit of replication here. However, it may be necessary to include several individuals in each container so that they can initiate reproduction for the experiment to work, even if those individuals are not at the scale at which replication is required. Similarly, we might be interested in how conspecific density affects growth: multiple individuals of the same species would be placed into a plot for one treatment, while single individuals would be placed in plots for the other treatment. In this example, we must necessarily include multiple individuals in some plots in order to create the treatment itself, but the plot is clearly the scale at which the treatment is applied, and hence the appropriate scale of baseline variation and replication.

Subsampling for improved precision and maximizing power

In the above examples, replication below the level at which the factor of interest is applied is a concession to practicality or necessity. Oftentimes, however, replicating at lower levels has a merit of its own: to improve the

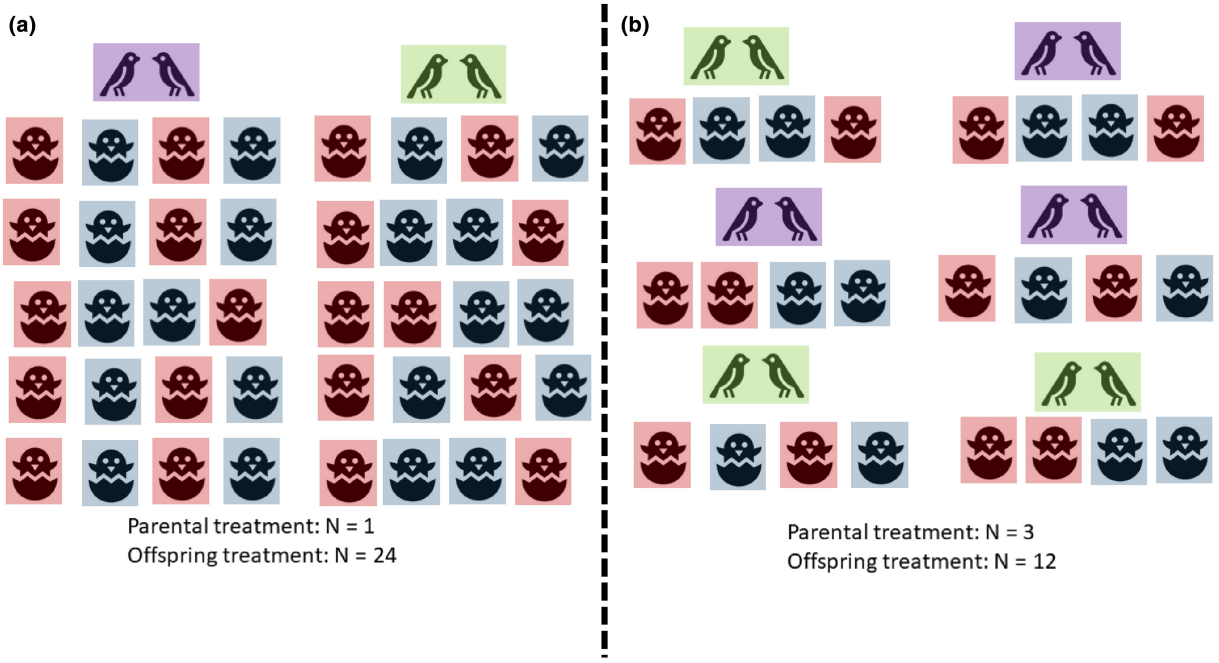


FIGURE 3 Two scenarios for a multilevel transgenerational experiment. Scenario (a) involves two sets of parents; each set is exposed to a different environment, and their offspring are exposed to two different environments later (different environment treatments are represented by different colours). In this design, the scale of replication for the parental generation is parents, so only one replicate has been included at this scale. No reasonable inferences about parental environmental effects driving any differences in offspring phenotype are possible here, but the effect of offspring environment on offspring phenotype can be assessed. Scenario (b) shows an experimental design where both the parental and offspring environments are replicated at the appropriate scales. Note that Scenario (a) involves more work in terms of measuring offspring phenotypes but is unreplicated at the parental level. This illustrates the point that inadequate designs can be just as laborious as good designs.

precision of the estimate of the response variable. Some biological traits or responses are easily captured by a single measurement. For example, measuring the maximum width of a barnacle is relatively straightforward, and measurement error will be trivial. Over the time course of a few minutes, that barnacle will not grow detectably, so there is no within-individual variation in that response variable over the measurement window. On the other hand, measuring that barnacle's metabolic rate will be much less straightforward—estimates of metabolic rate are likely to have more measurement error and show real temporal variation within any one individual. Hence, using a single measure of metabolic rate for a single individual could exaggerate the amount of variation that we would estimate within treatments (error, within- and among-individual variation would all be contributing to the estimate). Imagine a scenario where we ask whether barnacles growing on vertical settlement plates have different metabolic rates from barnacles growing on horizontal settlement plates—clearly, settlement plates are the scale at which the treatment is applied, so settlement plate is the appropriate unit of replication. But, given that metabolic rate is a noisy trait subject to within- and among-individual variation and nontrivial measurement error, we may benefit from subsampling—from measuring the metabolic rate of multiple individuals per plate and measuring the same individual multiple times. Subsampling in this case will greatly increase the precision of your estimate, yielding a better estimate of among-replicate variability with less contribution from measurement error and within-individual variability. Consequently, the power of your analysis is likely to go up.

I often use subsampling to increase precision in my experiments, but this requires additional effort. A key factor in deciding whether to subsample is the degree of within-sample variability (due to either measurement error or true within-replicate variability)—if variability is high, subsampling may be beneficial. But if subsampling is very costly or time-consuming, such that you must trade-off the number of subsamples that you can do directly with the number of replicates, then subsampling is harmful—instead, your effort should be devoted entirely to maximizing replication at the appropriate scale. In practice, however, subsamples can often be less costly or time-consuming to make, relative to replication at the appropriate scale. For example, it might be laborious to set up aquaria with various treatments but trivial to add a few extra snails to those aquaria; an analysis might benefit from multiple snails being measured within each (while ensuring that aquaria are also replicated; Figure 1c). Basically, when the costs of subsampling are relatively low and the possibility of within-replicate variability is high, then subsamples are probably worthwhile. In all other instances, it is either inefficient or pointless to replicate at lower scales.

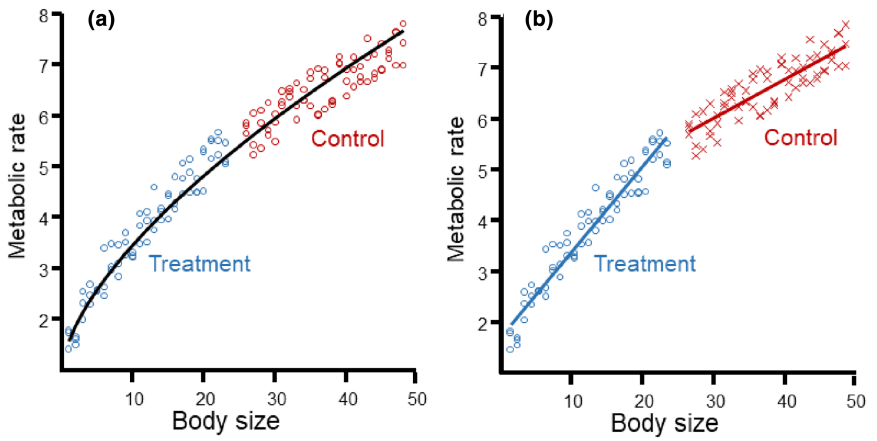
Replicating at higher scales

For completeness, I should mention that replication at higher scales than the level of the factor of interest is also possible. Replicating at this scale yields generality in either time or space, and can provide formal tests of how a factor interest might change across these. For example, repeating a predator-exclusion experiment at 10 locations enables you to ask whether there is a location by predator interaction, or, in other words, are the effects of predation consistent in space? But if you are interested in *why* predator effects differ between locations—you would need to replicate locations with and without your putative causative agent (Figure 2).

DISENTANGLED FACTORS AND CONFOUNDING

Most studies in ecology and evolution involve multiple factors of interest or a factor of interest and other experimental factors that may not be of specific interest, but are useful for including in order to reduce unexplained variation. For example, a study on the effects of a pharmaceutical by-product on the metabolic rate of tadpoles might also include body size in the analysis because body size has strong effects on metabolic rate. Whenever we have more than one factor that varies, we risk introducing ambiguous effects, whereby it becomes difficult or impossible to disentangle the influence of one factor relative to another. Below, I will explore some of the common issues and how to avoid them.

Including covariates can be a useful way of creating more sensitive analyses or examining how the effect of one factor varies with the value of another factor, but such approaches have important limitations that are often overlooked. Including body size when examining the effects of a pharmaceutical by-product on metabolism was a good example of using a covariate to enhance analytical power. But let's imagine that the pharmaceutical by-product of interest affected the growth rates of the tadpoles: controls grew much faster than those exposed to the pharmaceutical, such that, at the time of measuring metabolism, control tadpoles were larger than those in the treatment group, with no overlap in body sizes between the groups (Figure 4). We can analyse these data with body size as a covariate, but the analysis is unreliable because it implicitly assumes that the relationship between body size and metabolism remains exactly the same beyond the domain of body sizes that were measured in each treatment (Quinn & Keough, 2002). Such assumptions are unsupported and, in many instances, wrong. In a standard analysis of covariance, the model is assuming that the relationship between body size and metabolism remains linear beyond the range for which there is data. In this case, there is a (realistic) nonlinear relationship between body size and metabolism. Figure 4 illustrates how the model would inappropriately find an interaction between body size and the pharmaceutical effect where, in fact, there is no effect of anything other than body size. In short, covariates are not panaceas, and we need good representation of both factors of interest in combination if we are to reliably disentangle their relative effects. This also provides a good illustration of why it's essential to visualize your data before analysing it. Thus, it is important to have adequate overlap in the covariate of interest across different treatment groups when using any model with a mix of categorical and continuous factors. In some cases, this might



(c)

Factor	df	MS	F	p
Treatment	1	7	79	<0.001
Body Mass	1	91	978	<0.001
Interaction	1	12	136	<0.001
Error	134	0.09		

FIGURE 4 Two analytical approaches for the same simulated dataset. In scenario (a), the data show a nonlinear relationship between body size and metabolic rate, and there is no real effect of the treatment relative to the control on y; rather, it simply affects x. Scenario (b) [with the attendant ANCOVA table (c)] shows the dangers of analysing data where the covariate (body size) range does not overlap across the factors of interest. The analysis will indicate there are strong treatment effects on metabolic rate and that the effect interacts with body size, even though there is no effect at all. While the lack of overlap has been exaggerated here relative to real-world studies, it illustrates the nature of the problem. It is always worth visualizing data before conducting such analyses and, if necessary, restricting your analysis to the range of data where the treatment overlap with regards to the covariate.

require the extreme measure of ignoring datapoints that sit well outside the area of overlapping covariate ranges, but a better solution is to ensure adequate overlap of the covariate during the data collection phase.

Confounded factors

Biologists can recognize some types of confounding better than others. For example, imagine I fed cockroaches either high-protein food or low-protein food and kept all the high-protein individuals in 20°C chambers, and all the low-protein individuals in 25°C chambers. Clearly, any differences could be driven by protein or temperature.

But there are more subtle sources of confounding that slip their way into experimental designs. For example, imagine we are interested in the effect of the larval food regime on adult performance in the field. Some larvae receive lots of food, and other larvae receive less food, and when they metamorphose, they are deployed into the field. Because larvae that receive lots of food metamorphose much sooner, they might become adults after only 7 days, whereas larvae that receive less food metamorphose after 15 days. In this case, we have a problem because the field deployment date is confounded with the larval food regime—we cannot be sure whether any differences between our two larval food regimes are due to food, or because cohorts deployed on different dates will always have slightly different performance. This problem isn't avoided by keeping the high food treatments in the laboratory until the low food treatments also metamorphose, as then we would be confounding 'experience as adults' with larval food treatments. Instead, we must create a more complicated design where we establish low food treatments 8 days before the high food treatment, as well as another low food treatment on the same day as the high food treatment so as to disentangle the effects of adult deployment day from larval food environment effects. Ideally, we might embellish the experiment further to disentangle the effects of starting days, such that we have multiple start days, multiple end days and a good representation of high and low larval food regimes throughout these days (Allen & Marshall, 2010).

Collapsing replication

Sometimes good experimental designs go bad. In some instances, biologists can lose sight of what the unit of replication actually is, especially in convoluted designs, where good design principles might be adhered to in one phase of the experiment but are accidentally abandoned later in the experiment (Figure 5). For example, imagine you are interested in whether evolving in the presence of a predator affects the heavy metal tolerance of offspring in *Daphnia*. We might create five evolutionary lines that experience predators and five that do not—here 'line' is the unit of replication. After 10 generations of

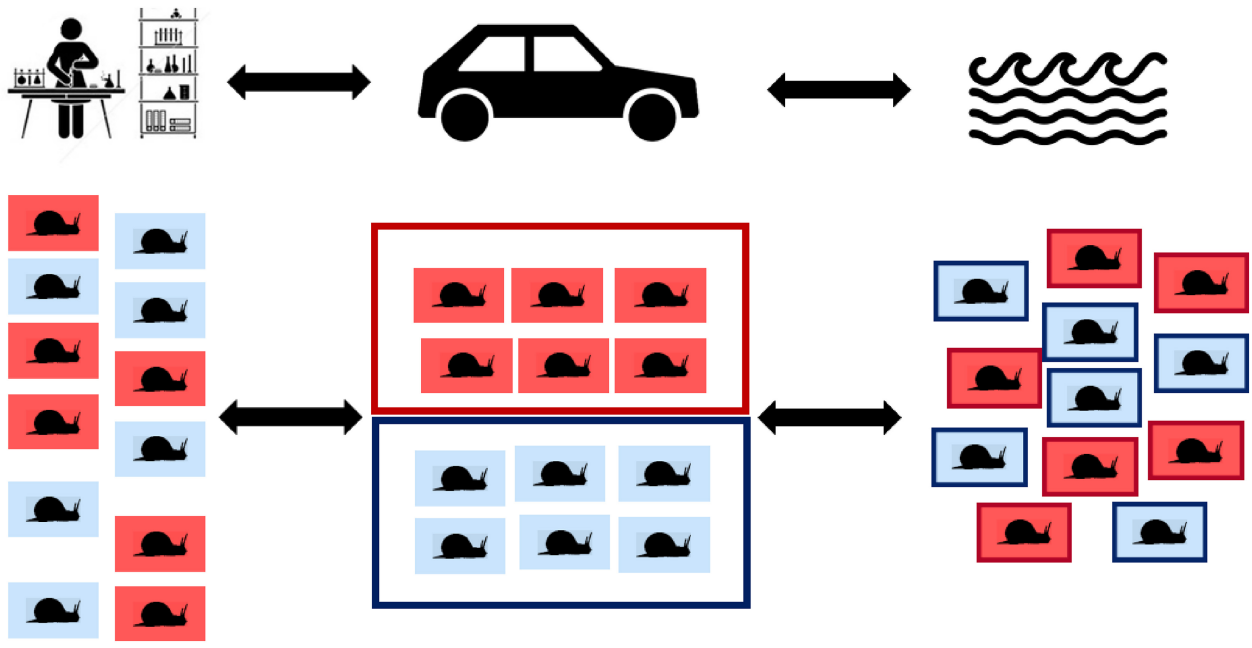


FIGURE 5 An example of 'collapsing replication', whereby a treatment is applied appropriately at the level of individuals, but then upon transporting the replicates either to or from the field (hence the bidirectional arrows), the treatments are segregated, and the treatment is now confounded with the transport container. Inferences about the effects of treatment will subsequently be much weaker, even though at either end of the process, the treatments were appropriately applied. At best, the estimate of the effects of interest will be noisier than they would otherwise be; at worst, all of the differences are due to transportation container effects.

experimental evolution, we might then take 50 offspring from each predator line, pool the samples into one holding container for the predator-exposed lines and one holding container for the predator-free lines, and then allocate those offspring to either a heavy metal treatment or a control vial. I call this mistake 'collapsing replication' whereby the integrity of the experimental replicates has been compromised because the replicates have collapsed into one confounded factor. We can no longer tell whether it was evolution in the presence of the predator or the holding container that drove any differences between groups.

Sometimes a good design can be corrupted at the last moment. Imagine a laboratory experiment where individuals are exposed to warm or cool temperatures, and we want to examine the effects of this exposure in the field, so we transport the individuals in insulated containers into the field (Figure 5). If logistical constraints mean that we transport all the warm individuals in one insulated container and all the cool individuals in another, we have now confounded the container with an experimental treatment. The same problem can occur in reverse, whereby field-collected individuals are transported back to the laboratory, where the field collection condition is confounded with the transport container or date of transport.

There are multiple advantages for avoiding this type of confounding. As I've noted, confounding will always reduce the precision of your estimate of the effect of interest. Given all the trouble we take to apply treatments carefully, and measure phenotypes accurately, it seems a shame to add imprecision unnecessarily through an inadequate design. More practically, anyone who has done experiments has experienced disaster—a dropped tray, a contaminated jar, a spilt vial. When all of a single treatment is clumped in space (e.g. container) or time, we concentrate all of our replicates for a particular treatment, which means that when disaster strikes, we are more likely to lose all of the replicates for that treatment, such that we have no replicates left for that treatment level. Interspersing replicates from different treatments therefore has the practical advantage of not putting one's eggs into a single basket, in addition to its experimental design benefits. Overall, I would advise to avoid systematic covariance between your factor of interest and other potentially confounding factors wherever possible, even if they seem minor.

CONCLUSIONS

My goal here has been to provide rules of thumb, examples and illustrations for how to identify the scale at which an inference is being made, or a treatment that is being applied. Of course, I could not include all possible examples, but I hope some principles have emerged that can be applied to any situation. Most of all, I encourage all researchers to think hard about scales of replication as empiricists, supervisors, reviewers and editors. I also think we need more conversations about experimental design—it is something most of us do, but it surprises me how little we talk about it openly. Most conversations about experimental design occur when it is inadequate or disqualifying, and these conversations often come too late. There is also tremendous scope for optimizing experimental design: once an adequate design with the minimum requirements has been identified, there are lots of additional opportunities for improving the efficiency, cost effectiveness, inferential power and generality of an experiment. My hope is that future conversations can focus on these more positive aspects, but we must first reduce the proportion of studies suffering from inadequate experimental design. Finally, to those trainee scientists who might be worried about their own designs: please use this piece as a catalyst for discussing your designs with your advisors, mentors and committees; do not assume that because they are more established, their intuition for experimental design is flawless.

ACKNOWLEDGEMENTS

I owe Mick Keough and Gerry Quinn a debt of gratitude for training me in experimental design; any failings in that regard are mine, not theirs. I am very grateful to Jon Chase for his insightful comments in revising this work. Bob Wong and members of the Marine Evolutionary Ecology Group provided helpful feedback on earlier versions of this manuscript. Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

No data in this paper.

ORCID

Dustin J. Marshall https://orcid.org/0000-0001-6651-6219

REFERENCES

Allen, R.M. & Marshall, D.J. (2010) The larval legacy: cascading effects of recruit phenotype on post-recruitment interactions. *Oikos*, 119, 1977–1983.

Heffner, R.A., Butler, M.J. & Reilly, C.K. (1996) Pseudoreplication revisited. *Ecology*, 77, 2558–2562.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54, 187–211.

Quinn, G.P. & Keough, M.J. (2002) *Experimental design and data analysis for biologists*. Cambridge, UK: Cambridge University Press.

Schank, J.C. & Koehnle, T.J. (2009) Pseudoreplication is a pseudoproblem. *Journal of Comparative Psychology*, 123, 421–433.