Trends in
Ecology & Evolution

CellPress

Review

Causal assumptions and causal inference in ecological experiments

Kaitlin Kimmel [1], Laura E. Dee, [2,*] Meghan L. Avolio, [1] and Paul J. Ferraro [3,4,*]

Causal inferences from experimental data are often justified based on treatment randomization. However, inferring causality from data also requires complementary causal assumptions, which have been formalized by scholars of causality but not widely discussed in ecology. While ecologists have recognized challenges to inferring causal relationships in experiments and developed solutions, they lack a general framework to identify and address them. We review four assumptions required to infer causality from experiments and provide design-based and statistically based solutions for when these assumptions are violated. We conclude that there is no clear demarcation between experimental and non-experimental designs. This insight can help ecologists design better experiments and remove barriers between experimental and observational scholarship in ecology.

Highlights

Causal inferences require causal assumptions. To formalize the assumptions required to draw causal inferences from experimental data, scholars have leveraged insights about causal inference in observational settings.

Even carefully designed experiments may face challenges in satisfying four important causal assumptions. Ecologists sometimes acknowledge and address these challenges but do not have a cohesive framework for understanding them.

When the validity of a causal assumption is questionable, ecologists can apply design-based and statistically based solutions.

Despite popular wisdom, no clear demarcation exists between experimental and non-experimental designs. When inferring causal relationships from data, experimentalists need to be just as careful as non-experimentalists in assessing the validity of their assumptions.

Experimentation in ecology

Experiments are the primary tool in ecology for quantifying causal relationships. Causal inference is facilitated by the experimenter's control over variation in the causal variable [i.e., the **treatment** (see Glossary)]. This control ensures that the cause precedes the effect and allows the experimenter to assume that the source of variation in the cause is not systematically related to variation in outcome, as it might be in observational designs.

Given the importance of experiments in ecology, ecologists have written extensively about experimental designs and analyses (e.g., [1–4]), including, for example, how replication and pseudoreplication can impact inferences (e.g., [5–8]) and how experimental treatments may not align with actual ecological phenomena (e.g., [9]). Further, ecologists widely acknowledge that experimental results may not generalize from one site to another, a limitation they have sought to overcome through widely distributed experimental networks (e.g., [10,11]).

Ecologists have paid less attention to the core assumptions required to infer causal relationships from correlations in experiments. Experimental data never 'speak' by themselves [12]. Only the combination of data and assumptions allows ecologists to infer causality from experimental data. Violations of these causal assumptions have important implications for what we can learn from an experiment. Given the complex realities of conducting experiments, most researchers are unlikely to implement the **ideal experiment** in which none of the core causal assumptions is violated.

Formalizing causality in ecology

Scholars who studied causality for decades have formalized core assumptions for inferring causality and developed solutions for the myriad contexts in which the assumptions may be violated [12–18]. These insights, however, have not been widely discussed in experimental ecology, where causal inferences are often not questioned (see [19–21] for discussions in observational

[1]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO, USA
[3]Carey Business School, Johns Hopkins University, Baltimore, MD, USA
[4]Department of Environmental Health and Engineering, a joint department of the Bloomberg School of Public Health and the Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

*Correspondence:
Laura.Dee@colorado.edu (L.E. Dee) and
pferraro@jhu.edu (P.J. Ferraro).

ecology). Without a deeper understanding of the core assumptions that make causal inferences from experiments possible, the causal conclusions ecologists draw from experiments may be erroneous.

Here, we describe the core causal assumptions and describe solutions that are available when one or more assumptions are violated. Our goal is to help researchers make the most of their experimental data and better evaluate the credibility of causal inferences in experimental studies.

Causality and potential outcomes perspective

Ecologists are familiar with good experimental design practices like randomizing treatments and having multiple replicates. However, even well-designed experiments rely on assumptions that, when left unexamined, can lead to inaccurate estimates of the **causal effect** of interest. Before exploring these assumptions, we introduce terms and concepts of causality using the **potential outcomes** perspective (the Neyman–Rubin Causal Model) [18,22,23]. We illustrate this perspective and its notation through a hypothetical example of using exclosures to remove herbivores.

Say we are interested in the causal effect of eliminating herbivores on species richness on the *i*th plot (Figure 1). In this example, the causal variable can take only two values: a herbivory **control** ($T_i = 0$, where T corresponds to the treatment) or a no-herbivory treatment that uses a herbivore exclosure ($T_i = 1$). When the plot is a control, its species richness outcome is $Y_i(0)$. When the same plot is treated, its outcome is $Y_i(1)$. $Y_i(1)$ and $Y_i(0)$ are called potential outcomes because both are potentially observable. The difference between these potential outcomes is the plot-level causal effect of eliminating herbivory.

In other words, a causal effect is defined as the difference in outcomes between two states of the world. The challenge is that, in practice, only one of these outcomes can be observed at a point in space and time. We can only observe $Y_i(1)$ on a treated study **unit** and $Y_i(0)$ on a control study unit. The unobserved outcome for the study unit is a **counterfactual outcome**. Our inability to observe the same study unit under both treated and control conditions is the Fundamental Problem of Causal Inference [17,18,24] and implies that we cannot estimate unit-level treatment effects.

Although we cannot estimate a treatment effect for each study unit, we can estimate an **average treatment effect (ATE)** across study units. When treatment assignment is randomized, the treated and control units do not systematically differ. Thus, they have the same expected potential outcomes. We can then assume that the expected $Y_i(0)$ on control units is equal to the expected unobserved (counterfactual) $Y_i(0)$ on treated units, and the expected $Y_i(1)$ on treated units represents the counterfactual expected $Y_i(1)$ on control units. With treatment randomization in the herbivory experiment (Figure 1), we can estimate the expected value of $Y_i(1)$ for all plots using the average outcome in the treated plots and the expected value of $Y_i(0)$ for all plots using the average outcome in the control units. Therefore, randomization allows one to use the observable data from the experiment to estimate the ATE across all units:

$$\frac{1}{N}\sum_{i=1}^{N}[Y_i(1)] - \frac{1}{N}\sum_{i=1}^{N}[Y_i(0)]. \quad [1]$$

With random assignment of the treatment, we can estimate, without **statistical bias**, the average causal effect simply by taking the difference in observed outcomes between the treatment and control units (Equation 1).

Glossary

Average treatment effect (ATE): the expected change in a randomly selected unit from the target population when the unit moves from one treatment condition to another; that is, how the outcome would change, on average, if all units moved from one treatment condition to another.
Causal effect: a comparison of the potential outcomes under two values of a treatment.
Compliers: units that receive a specific value of a treatment when assigned that value and not otherwise (the units 'comply' with their treatment assignment).
Control: a term that can describe the absence of treatment (untreated) or can describe a baseline condition (e.g., species richness = 1).
Counterfactual outcome: the outcome that a unit would have experienced if the unit had received a different treatment (i.e., 'contrary to fact').
Excludability: outcomes respond solely to a treatment itself and not to another causal pathway that is set in motion by the assignment of a treatment.
Falsification tests: statistical tests that leverage theory to provide evidence that core causal assumptions may be invalid.
Ideal experiment: an experiment whose design does not violate core causal assumptions described in this review.
Interference: when the potential outcomes of one experimental unit depends not only on its own treatment status but also on the treatment status of other units; in other words, when the treatment status of one unit affects the outcomes of other units. 'No interference' is one part of what statisticians call the 'stable unit treatment value assumption' (SUTVA).
Internal validity: in a particular context, the degree to which the evidence supports a causal claim (i.e., the degree to which rival, noncausal explanations can be eliminated).
Measurement error: error in the measurement of outcomes or treatment status.
Multiple versions of treatment: each treatment condition has more than one version and thus each unit may have more than one potential outcome per treatment condition. 'No multiple versions of treatment' (a.k.a. 'no hidden treatments') is one part of what statisticians call the SUTVA.

Core assumptions in causal inference

Randomization alone does not guarantee that we can estimate an accurate casual effect from Equation 1. To estimate an average causal effect without bias requires that the experimental design also satisfies four core assumptions (Figure 2): **excludability**, no **interference**, no **multiple versions of treatments**, and no **noncompliance**. However, even the most carefully planned experiments may violate one or more of these assumptions because of logistical constraints or unforeseen field or laboratory conditions. Here, we describe the four assumptions and explore how they may be violated in ecological experiments (see Boxes 1 and 2 for more examples). We then provide solutions to avoid these violations and address them when they occur (Table 1).

Noncompliance: when some units do not receive their assigned treatment.
Partial identification: a statistical approach to test how sensitive a causal claim is to causal assumptions; the approach yields estimated bounds on the causal effect rather than a point estimate.
Potential outcome: the value of a unit's outcome under a particular value of a treatment.
Rival pathway: an alternative, noncausal pathway that connects a treatment and outcome.
Statistical bias: a systematic difference between the true value of a causal effect and the results from an estimation procedure. In contrast to sampling variability, bias does not decline with more data.
Treatment: synonym for 'cause' or 'causal variable'; often thought of as binary (e.g., treatment and control) but can be multivalued (e.g., species richness can be 1, 2, 3, …); can be thought of as an intervention or a manipulation of an attribute of a system.
Unit: the animal, plant, place, or thing that is exposed to a treatment at a particular time. Note that a single animal, plant, place, or thing at two different times comprises two different units.

Excludability: potential violations

To infer a causal effect from data generated in a randomized experiment, one must assume that excludability is satisfied, meaning that the process by which treatments are assigned has no effect on potential outcomes except through its effect on a unit's treatment status [15]. With this assumption, we can justify excluding the treatment assignment method when calculating the ATE (Equation 1). However, ecological processes cannot be directly changed. Instead, experimentalists manipulate ecological processes through laboratory or field interventions. For example, experimenters cannot directly change precipitation but instead rely on methods like shelters (Figure 2A). Treatment assignment methods like shelters are therefore part of the causal pathway being studied. We generally ignore their role because we assume that there is no connection between the method and the outcome except through the treatment itself (i.e., no red line in Figure 2A). However, shelters may change other aspects of the ecosystem that impact the outcome of interest.

If this assumption is violated, researchers cannot rule out rival explanations for the data patterns they observe. For example, in many biodiversity experiments that quantify the effect of species richness on productivity, researchers maintain the original richness treatment by weeding out species that enter the plot. Weeding intensity is likely to be correlated with the diversity treatment because plots with higher diversity tend to have fewer invaders than lower-diversity plots [25,26]. Weeding may increase the productivity of low-diversity plots by aerating the soils [27,28]. Therefore, a researcher cannot be sure whether differences in outcomes (productivity) are caused only by the diversity treatments or whether weeding to maintain the treatments also had an effect.

Excludability violations can also arise when **measurement error** systematically changes with treatment status [29]. For example, in herbivory experiments, plants in plots with herbivores are likely to be trampled and eaten. Identifying plants may be more difficult in these plots than those without herbivores. A correlation between the treatment status and measurement error creates an excludability violation.

Excludability: solutions

Addressing potential excludability violations is best done at the design stage. At this stage, one can determine which way to implement the treatment that would be least prone to creating unintended causes. For example, one can anticipate when measurement error may be correlated with treatments and develop measurement protocols to mitigate these threats.

Even under the best designs, excludability may be violated. To address potential violations, two categories of statistical approaches are available, although none can confirm that excludability is satisfied (i.e., none can disconfirm all **rival pathways** from treatment assignment to outcomes).

First, one can use theory and field knowledge to identify how excludability may be violated (Figure 2A, red line) and then either control for them statistically or detect their influence on the outcome

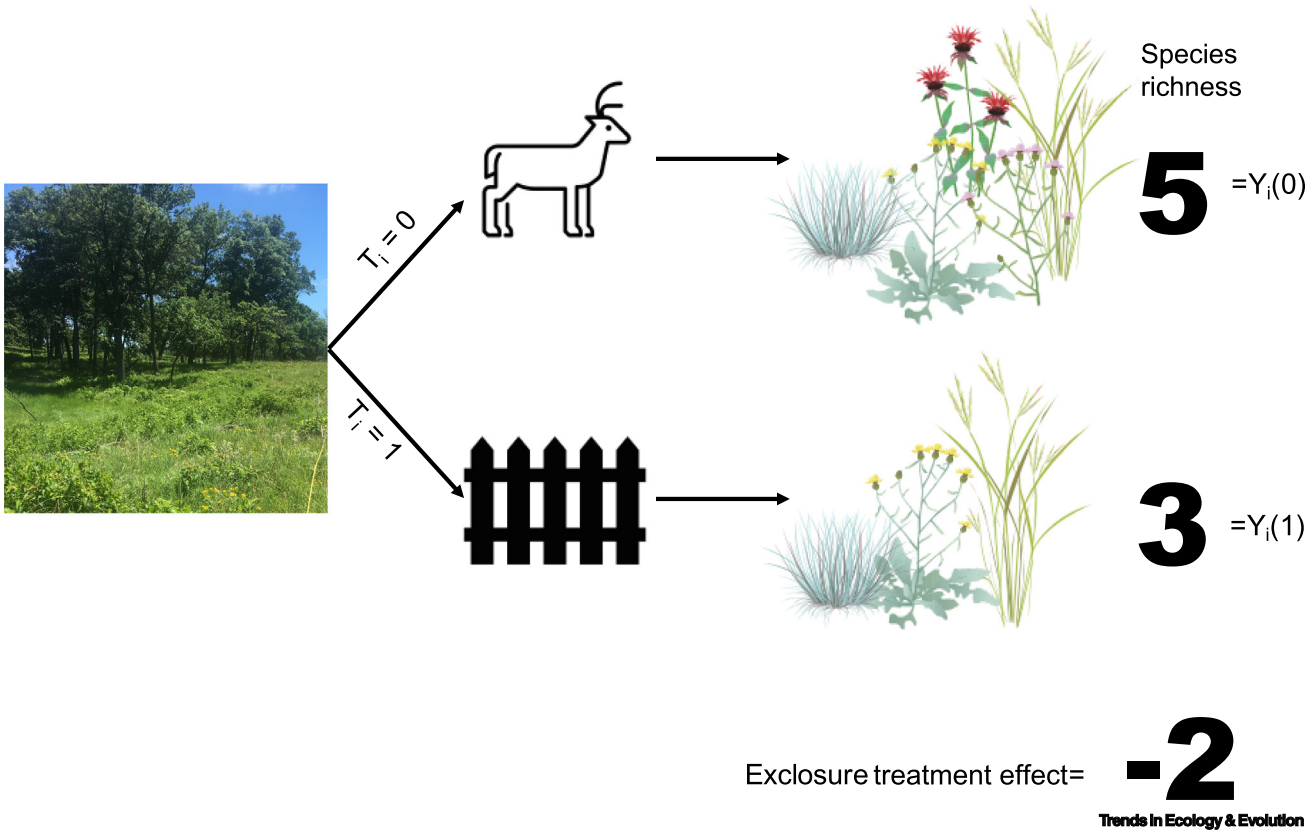

Figure 1. In the potential outcomes perspective, every study unit has a potential outcome for each treatment of the experiment. Here, we illustrate an example of an experiment that quantifies the effect of herbivore removal on species richness. The two treatments for the *i*th plot are denoted by T_i and the potential outcomes by Y_i. Here, there are two treatments: $T_i = 0$ is the control where herbivores can graze and $T_i = 1$ is the treatment where exclosures prevent herbivores from grazing. If the plot is a control, it has a potential outcome of five species. If it is treated, it has a potential outcome of three species. The difference between $Y_i(1)$ and $Y_i(0)$ is the unit-level causal effect of the exclosure treatment. Given that researchers cannot observe both $Y_i(0)$ and $Y_i(1)$ for the same plot, they must make assumptions about how randomized treatment conditions represent counterfactual conditions, which allows researchers to estimate an average treatment effect. Plant species images by Tracy Saxby, IAN Image Library (https://ian.umces.edu/imagelibrary).

via a **falsification test**. To control them statistically, one could try to measure the source of the violation (e.g., soil aeration) and control for it directly by including it as a covariate in a regression, similar to what is done in observational studies [13–16,30]. Alternatively, when the potential sources of an excludability violation are known but not observable and therefore not amenable to statistical control, a falsification test can be used (a.k.a. a placebo test, placebo design, or test of known effect [16,30]). A falsification test uses either a placebo treatment or a placebo outcome. A placebo treatment is a treatment that impacts the outcome through rival pathways only. For example, in drought experiments (Figure 2), ecologists set up shelter controls that do not change precipitation but induce other causal pathways from shelters to the outcome should they exist [31]. A placebo outcome is an outcome that could not be affected by the treatment but could be affected by rival pathways created by the treatment assignment method. In both cases, if the estimated placebo effect is statistically different from zero at an ecologically meaningful magnitude, there is reason to believe that the experimental design suffers from an excludability violation.

Second, one can use theory and field knowledge about the potential excludability violation to either bound the value of the target treatment effect using **partial identification** or assess how

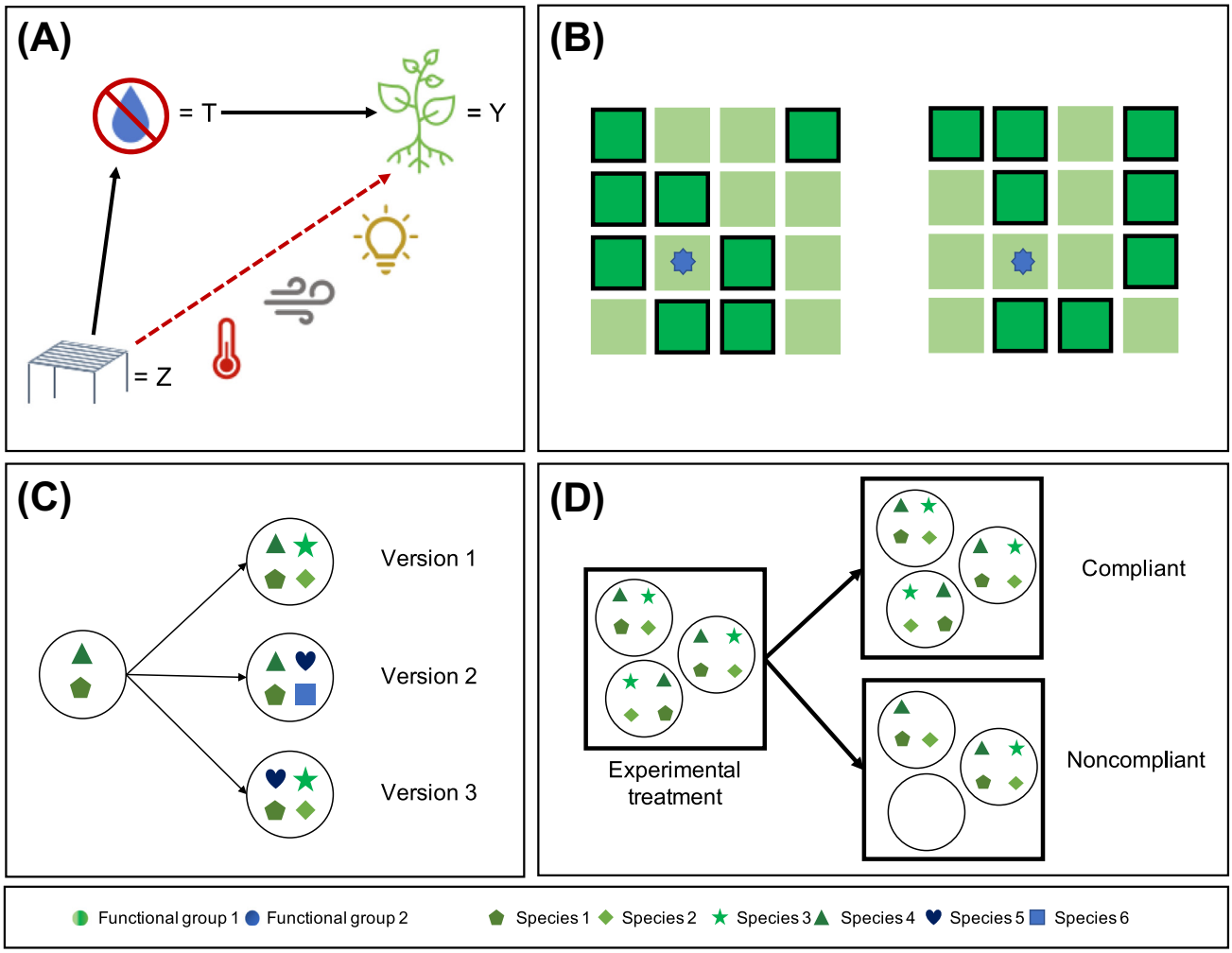

Figure 2. Illustrations of (A) violation of excludability, (B) interference, (C) multiple versions of treatments, and (D) noncompliance. In (A), we show an excludability violation in a drought experiment. The drought treatment ($T = 1$) affects plant growth [$Y(1)$], a causal effect represented by the black line between the droplet and the plant. The drought treatment is applied using a shelter [$Z(1)$] represented by the black line between the shelter and the droplet. The shelter also impacts plant growth by altering the temperature, humidity, and light, represented by the red line. Thus, the outcome is affected by the treatment and by the technique used to manipulate the treatment. In (B), we illustrate interference in two potential arrangements of a herbivore exclusion experiment. Plots (squares) with black borders are fenced to exclude herbivores; other plots serve as controls. The starred plot has different potential outcomes depending on the treatment plot arrangement. In (C), we illustrate multiple versions of treatments in a biodiversity experiment. The potential productivity, for example, with two species and the potential productivity with four species may depend on the species' identity. Here, we show multiple versions of the four-species treatment: each version can lead to a different potential outcome. In (D), we illustrate noncompliance of an experimental treatment where four species are planted. A plot complies with its treatment when it has all four planted species growing in it. A plot does not comply when it has a different number (or identities) of species growing in it.

sensitive the estimated effect is to a potential violation of excludability (sensitivity tests to hidden bias). For example, in a study that estimated how cash transfers to poor Indonesians affected tropical forest loss [32], the authors argued that the estimated negative treatment effect was not sensitive to an excludability assumption violation. To drive the estimated effect to zero, one would have to assume there was no measurement error in the forest loss data and the unobserved confounding variable(s) explained more than half of the variation in forest loss and was one-fifth as important in affecting the variation in the treatment variable as were the control variables already in the analysis. The authors argued that this large degree of confounding was unlikely. For more details, see [33,34] on partial identification and [16,32] on sensitivity tests.

Box 1. Examples from abiotic manipulation experiments

Here we explore potential violations in each assumption required to infer causality in abiotic manipulation experiments. These experiments include manipulations of nutrients, temperature, and rainfall.

Excludability

Passive warming experiments use cone-shaped chambers to increase temperature. The walls of the chambers can also affect airflow, light conditions, and dispersal of other species into plots [54] and thus affect the outcomes of interest like diversity, productivity, and stability.

Interference

In nutrient addition experiments that aim to estimate the effect of nutrients on species diversity and composition [55,56], the results can be challenging to interpret when the diversity and composition of one plot affects the diversity and composition of other plots through, for example, dispersal (e.g., P. Hawthorne, PhD thesis, University of Minnesota, 2012; https://www.cedarcreek.umn.edu/biblio/fulltext/Hawthorne_umn_0130E_12586.pdf). In this case, the causal effect from nutrient addition on a plot depends on whether the neighboring plots were treated or control plots.

Multiple versions of treatments

In nitrogen addition experiments that aim to estimate the effect of increased nitrogen on ecosystem functions, the results can be challenging to interpret when the effect depends on the different forms of nitrogen (e.g., urea vs ammonium nitrate) or different manufacturers of a particular form, which may vary in their potency, residency time, and uptake into plants. If the treatment effect on a plot depends on which version of nitrogen it received, drawing clear inferences from an experiment can be challenging if multiple versions of nitrogen were used across space or time.

Noncompliance

- Nutrient addition experiments in arid systems may fail to deliver the treatment when no rain falls (fertilizers need sufficient water to enter the soil). When lots of rain falls in mesic systems, nutrients may just run through the soil. In both cases, treated plots are noncompliant.
- Nutrient addition experiments may accidently deliver additional nutrients to control plots when nutrients added to one plot leach into other plots. In this case, control plots are noncompliant. Ecologist often avoid this spillover by sufficiently spacing their plots when space allows.
- In exclusion experiments, the fences or cages may only partially exclude the target species. In other words, the species that the fences or cages attempt to exclude may still enter the plots. Thus, these plots are noncompliant because some herbivory is still occurring.

No interference: potential violations

Another core assumption of causal inference is that there is no interference between units, meaning that the outcome of a unit depends only on its own treatment status but not on the treatment status of another unit [35]. If each treatment assignment configuration creates a different set of potential outcomes, there is interference between units. Interference is not the same as spatial correlation (e.g., [3,36–38]), something for which some ecological experiments already account. Plots can be nested within a site and not interfere with each other.

To illustrate, consider a randomized experiment that uses fences to exclude herbivores from plots to quantify the causal effect of herbivores on carbon sequestration. Exclusion in some plots can have effects on outcomes in unexcluded plots. First, many excluded plots around an unexcluded plot may reduce the interest of herbivores in the unexcluded plot (Figure 2B, right panel), increasing carbon sequestration in this unexcluded plot compared with what would have occurred had it been surrounded by fewer excluded plots. Alternatively, some, but not many, excluded plots around an unexcluded plot may displace herbivory pressure onto other unexcluded plots (Figure 2B, left panel), reducing carbon sequestration in the unexcluded plot. The average causal effect of herbivory depends on the specific treatment assignment configuration. This dependence is an issue because control units are no longer unaffected by the treatment, which can lead to biases when estimating causal effects.

Box 2. Examples from biotic manipulation experiments

Here we explore potential violations in each assumption required to infer causality in biotic manipulation experiments. These experiments include manipulations of richness, composition, and other attributes of community structure, along with manipulations of trophic interactions.

Excludability violations

- Species removal experiments in rocky intertidal ecosystems often scrape sessile invertebrates (e.g., barnacles) or macrophytes from substrates [57,58]. Scraping could lead to the removal of non-target species (e.g., snails, algae, other sessile invertebrates), which could affect outcomes like productivity or interactions of non-removed species.
- Exclusion experiments in aquatic environments use cages to prevent consumers from entering plots. The cages can also modify the abiotic environment (by changing wave energy in the rocky intertidal) and the biotic environment (by affecting species interactions, as in [59]), changes that can subsequently affect outcomes like survival, growth, and competition.

Interference

In biodiversity manipulation experiments that aim to estimate how changes in biodiversity effect ecosystem functions, the manipulation of biodiversity in one plot may affect the ecosystem functions in other plots via, for example, dispersal. A low-diversity plot with high-diversity neighbors may have different productivity than the same plot with low-diversity neighbors. In this case, the treatment effect from a change in biodiversity on a plot depends on the treatment status of the neighboring plots.

Multiple versions of treatments

In species-richness manipulation experiments that aim to estimate the effect of changes in the number of species on ecosystem function, the results can be challenging to interpret when the effect depends on the identity of the species. For example, imagine a simple experiment where plots are randomly assigned to be planted with either two species or four species chosen from a subset of species that grow at the site. The potential productivity with two species and the potential productivity with four species may depend on the identity of the species; for example, whether the species are rare or dominant species [60] or whether they belong to a particular functional group. In this case, the estimated effect of going from two to four species may depend on which subset of species is selected from all of the species that grow at a site and whether all possible combinations of two and four species are planted.

Noncompliance

In species-removal experiments in rocky intertidal ecosystems, researchers scrape sessile invertebrates (e.g., mussels, barnacles) from substrates. However, recolonization may occur. Thus, some locations that are assigned to the 'species-removal' treatment will instead have sessile invertebrates on the substrate. To address this noncompliance, ecologists often repeatedly scrape to minimize recolonization [61,62], but this can lead to excludability violations (see above).

No interference: solutions

Like actions to address potential excludability violations, actions to address potential interference are best done at the design stage. At this stage, one can identify ways that interference may arise and take steps to mitigate or eliminate its effects. In ecological systems, many forms of interference are local: the closer two units are to each other, the more likely it is that they will interfere with each other. Thus, one can either ensure that study units are spaced far apart (e.g., [39]) – thereby eliminating interference – or randomize treatments at larger spatial scales, such as groups of neighboring plots rather than individual plots, thereby containing the interference within the randomized units.

Another approach is to design the experiment to detect interference and adjust the estimated causal effect to incorporate interference. These designs, called randomized saturation designs [40], vary the treatment in ways that allow one to measure interference. For example, instead of randomizing half of all plots in each study block to the treatment group, one would vary the percentage of treated plots in each block (e.g., some blocks would have 25% of their plots treated, some would have 50%, and some would have 75%). With this design, one can quantify the effect that arises when the treatment status of a plot's neighbors affects a plot's outcomes.

Table 1. Design-based and analysis-based solutions to violations in core causal assumptions

Assumption	How to avoid or address violations in the assumption	Refs
Excludability	**Design stage** Experimental design: identify the ways in which the treatment can be applied and use the one least likely to lead to excludability Experimental design: develop measurement protocols to mitigate the potential for measurement error to be correlated with treatment conditions	
	Analysis stage Statistical control: measure sources of violation and control via covariates in a statistical model	
	Falsification tests: probe assumptions of excludability and seek evidence that assumptions may be invalid	[16,30,52]
	Partial identification: construct bounds on estimated causal effect based on assumptions about excludability violations	[33,34]
	Sensitivity tests: test how sensitive the estimated effect size or statistical significance is to potential excludability violations	[16,32]
No interference	**Design stage** Experimental design: space units sufficiently far apart	
	Experimental design: randomize treatments at larger spatial scales (e.g., randomize at blocks rather than plots)	
	Experimental design: randomized saturation design	[40,53]
	Analysis stage Redefine causal effect: acknowledge that estimated effect is conditional on treatment assignment arrangement	[41]
	Modeling adjustment: model the interference to quantify or remove its effects	[41]
	Partial identification: construct bounds on estimated causal effect based on assumptions about form of interference	[33,34,43]
No multiple versions of treatments	**Design stage** Experimental design: define each version as a treatment	[45]
	Experimental design: restrict treatments to a subset of versions	[45]
	Experimental design: randomize versions and take average across versions	[45]
	Analysis stage Redefine causal effect: acknowledge that estimated effect is conditional on an unknown distribution of versions	[45]
No noncompliance	**Design stage** Experimental design: design post-randomization procedures to ensure compliance (e.g., calibrate machinery, replant species, remove unwanted species)	
	Analysis stage Redefine causal effect: estimate the Intent to Treat effect	[15,48]
	Redefine causal effect: estimate the Complier Average Causal Effect	[15,49,50]

If interference cannot be fully addressed at the design phase, there are two main approaches to address it in the analysis stage. First, one can redefine the causal effect as the effect of the treatment conditional on the treatment assignment in the experiment. This shift in interpretation acknowledges that the causal effect of the treatment may differ with different treatment assignment patterns. Second, one can take approaches similar those described previously for addressing excludability violations. For example, one can try to model interference (e.g., by assuming it occurs along a weighted distance gradient from each unit) to adjust the causal estimate [41,42]. Alternatively, one could make assumptions about the spatial structure of interference to create bounds around the true ATE instead of reporting a point estimate only (i.e., partial identification) [33,43].

No multiple versions of treatments: potential violations

To make clear causal claims in experiments, there cannot be multiple versions of the treatments, meaning that a treatment must be consistent among all treated units [13,35]. Like interference, multiple versions of treatments poses a challenge for the interpretation of results from experiments because there are multiple potential outcomes per treatment status (Figure 2C). Each version may have a different causal effect. For example, consider a nitrogen addition experiment that used ammonium nitrate for the first several years then switched to urea in subsequent years. Whether the differences between these two fertilizer versions are consequential depends on whether the versions have different effects on the mechanisms that mediate the treatment effect [44].

No multiple versions of treatments: solutions

To address the threat to inference from multiple versions, there are three design-based approaches. First, one can use each version as a different treatment and then estimate the effect of each treatment version on the outcome. That approach, however, may often be logistically infeasible. The second approach, and the one likely to be the most popular in ecology, is to restrict the experiment to a single version or a small set of versions. Such experiments may have high **internal validity**, but their results may not generalize to other treatment versions if those versions affect the outcomes differently [44]. A third approach seeks greater generalizability by simply averaging over the versions, at the cost of potentially missing important ecological insights about mechanisms by ignoring differences across versions. In this approach, the researcher randomly assigns treatment versions that are a random draw from a distribution of versions (e.g., as the versions are distributed in nature or a different distribution, like a uniform distribution; see example in Box 2). The researcher would then interpret the estimated average causal effect as the expected treatment effect from a randomly drawn version that is assigned to a randomly drawn unit from the target population. For more methods to statistically address the threat of multiple treatment versions, see [45].

No noncompliance: potential violations

Causal inference requires that units receive the treatment they were assigned. If a unit's actual condition differs from its assigned treatment condition, there is noncompliance (Figure 2D). In ecological experiments where the treatment is applied by the researcher, noncompliance may be minimal, but it is possible in some study designs. For example, in species-richness manipulation experiments, researchers assign each plot a specific number of species. However, after a plot is planted, some of the species may not establish or persist because of, for example, successional dynamics and competition. Thus, some plots will have a realized richness that is lower than their assigned richness (e.g., [46,47]).

No noncompliance: solutions

Like actions to address excludability and interference, actions to address noncompliance are best done at the design stage. These actions include careful planning to ensure that deviations from the experimental protocol do not arise in the field, followed by careful monitoring to ensure, for example, that exclusion barriers are well maintained or that free-air CO_2 enrichment machines are well calibrated. However, actions to mitigate noncompliance can often lead to excludability violations. For example, weeding or re-seeding plots to maintain planted richness over time may affect ecosystem functions through causal paths that do not pass through the richness treatment. Frequent visits to plots to monitor equipment can cause soil compaction or other disturbances that can subsequently affect the outcomes that are the focus of the study.

When it is not possible to maintain 100% compliance, one can also address noncompliance in the analysis phase. Simply excluding noncompliant units from the analysis or redefining their treatment status based on the treatment they actually received is inappropriate unless one is willing to assume that noncompliant units are a random draw from the target population (i.e., noncompliance is unrelated to potential outcomes). When noncompliance is not random, as is likely for most ecological contexts, dropping units or redefining their treatment status introduces excludability violations: the treated and control units (or units across multiple treatment conditions) no longer have equal expected potential outcomes in the presence and absence of the treatment (i.e., randomization no longer guarantees unbiased inferences from the data).

Instead of discarding or relabeling noncompliant units, researchers can change the causal effect of interest and thus the interpretation of the estimate. Two popular approaches are to estimate the intent-to-treat (ITT) effect [48] or the **complier** average causal effect [49] (CACE) [or a Local Average Treatment Effect (LATE)] [50]. For example, a researcher may want to use Equation 1 to estimate the ATE, which is the expected effect of treatment exposure on a randomly chosen unit from the target population. With noncompliance, however, a researcher cannot estimate the ATE without making strong, untestable assumptions about the causes of noncompliance. Treatment randomization still allows the researcher to estimate the ITT effect, which is the expected effect of treatment assignment on a randomly chosen unit from the target population. Instead of asking whether the treatment causes an effect, one asks whether the attempt to manipulate the treatment variable has a causal effect (i.e., researchers analyze the units 'as assigned' rather than 'as treated'). In some contexts, like ecosystem management where the treatment 'as assigned' is of interest (e.g., restoration), the ITT may be a useful ecological parameter. In cases where it is not useful, researchers may be able to estimate the CACE, which is the expected effect of treatment for a randomly chosen unit that complies with its assigned treatment. Compliers are a subgroup of the experimental units and may not be representative of all units. Thus, the CACE may not capture a causal effect of scientific interest. However, it may be preferable to the ITT or a biased estimate of the ATE. For details on how to estimate the CACE through what are often called randomized encouragement designs or instrumental variable designs, see [51].

Concluding remarks

The potential outcome perspective helps to clarify, regardless of the empirical design, the causal effects that ecologists are trying to understand and the plausibility of the assumptions that are required to infer causality from correlations in the data. As noted by Pearl [12], data never speak by themselves, whether they come from an experiment or an observational study. They only speak when combined with untested – and often untestable – assumptions.

All empirical designs are judged by the extent to which they may deviate from the ideal experiment for the target research question, an experiment for which all assumptions described herein hold with certainty and thus an experiment unlikely to exist in reality. The goal of researchers is to minimize the deviations from this ideal through changes in design, changes in analyses, and changes in the interpretation of their real-world studies. From this perspective, one can see that, despite the popular wisdom, there is no clear demarcating line between experimental and observational studies. Broader understanding of this insight in ecology promises not only to improve empirical science in ecology (see Outstanding questions) but also to break down the unproductive, historical barriers between experimental and observational approaches in ecology.

Outstanding questions

How common are violations of the four assumptions in ecological experiments and are the violations serious enough to change conclusions about ecological processes?

Are certain types of ecological experiments more prone to certain violations?

How should we change the interpretation of causal effects from ecological experiments when interference and noncompliance are suspected?

When is variation in the treatment ecologically consequential? Threats to causal inference from multiple versions of the treatment are challenging to address because we need to know what variations in the treatment matter.

How can ecologists make inferences about intermediate causes when they are not randomized? Intermediate causes are on the path between treatment and outcome (often called 'mechanisms' in the causal inference literature). Ecologists may be tempted to make causal claims about these intermediate causes from experiments where only the treatment is randomized, not the intermediate cause.

How might we create incentives and norms for reporting on the validity of the four assumptions described in this review in ecological publications?

Declaration of interests

No interests are declared.

References

1. Gotelli, N.J. and Ellison, A.M. (2012) A bestiary of experimental & sampling designs. In *A Primer of Ecological Statistics* (2nd edn), pp. 163–206, Sinauer Associates
2. Hector, A. (2015) The New Statistics with R: An Introduction for Biologists. Oxford University Press
3. Steel, E.A. *et al.* (2013) Applied statistics in ecology: common pitfalls and simple solutions. *Ecosphere* 4, 1–13
4. Underwood, A.J. (1996) Experiments in Ecology (1st edn), Cambridge University Press
5. Kreyling, J. *et al.* (2018) To replicate, or not to replicate – that is the question: how to tackle nonlinear responses in ecological experiments. *Ecol. Lett.* 21, 1629–1638
6. Oksanen, L. (2001) Logic of experiments in ecology: is pseudoreplication a pseudoissue? *Oikos* 94, 27–38
7. Colegrave, N. and Ruxton, G.D. (2018) Using biological insight and pragmatism when thinking about pseudoreplication. *Trends Ecol. Evol.* 33, 28–35
8. Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54, 187–211
9. Korell, L. *et al.* (2019) We need more realistic climate change experiments for understanding ecosystems of the future. *Glob. Chang. Biol.* 26, 325–327
10. Borer, E. *et al.* (2017) A decade of insights into grassland ecosystem responses to global environmental change. *Nat. Ecol. Evol.* 1, 0118
11. Borer, E.T. *et al.* (2014) Finding generality in ecology: a model for globally distributed experiments. *Methods Ecol. Evol.* 5, 65–73
12. Pearl, J. (2009) Causal inference in statistics: an overview. *Stat. Surv.* 3, 96–146
13. Morgan, S.L. and Winship, C. (2014) Counterfactuals and Causal Inference: Methods and Principles for Social Research (2nd edn), Cambridge University Press
14. Imbens, G.W. and Rubin, D.B. (2015) Causal Inference: For Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge University Press
15. Gerber, A.S. and Green, D.P. (2012) Field Experiments: Design, Analysis, and Interpretation. W.W. Norton & Company
16. Rosenbaum, P. (2010) Design of Observational Studies. Springer
17. Rubin, D.B. (2005) Causal inference using potential outcomes. *J. Am. Stat. Assoc.* 100, 322–331
18. Rubin, D.B. (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701
19. Larsen, A.E. *et al.* (2019) Causal analysis in control–impact ecological studies with observational data. *Methods Ecol. Evol.* 10, 924–934
20. Grace, J.B. and Irvine, K.M. (2020) Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology* 101, e02962
21. Wauchope, H.S. *et al.* (2021) Evaluating impact using time-series data. *Trends Ecol. Evol.* 36, 196–205
22. Fisher, R.A. (1935) The logic of inductive inference. *J. R. Stat. Soc.* 98, 39–82
23. Neyman, J. (1990) On the application of probability theory to agricultural experiments. Essay on principles. *Stat. Sci.* 5, 465–472
24. Holland, P.W. (1986) Statistics and causal inference. *J. Am. Stat. Assoc.* 81, 945–960
25. Isbell, F.I. *et al.* (2017) Benefits of increasing plant diversity in sustainable agroecosystems. *J. Ecol.* 105, 871–879
26. Steinauer, K. *et al.* (2016) Convergence of soil microbial properties after plant colonization of an experimental plant diversity gradient. *BMC Ecol.* 16, 19
27. Currie, J.A. (1962) The importance of aeration in providing the right conditions for plant growth. *J. Sci. Food Agric.* 13, 380–385
28. Huang, B. *et al.* (1998) Effects of high temperature and poor soil aeration on root growth and viability of creeping bentgrass. *Crop Sci.* 38, 1618–1622
29. Millimet, D.L. (2011) The elephant in the corner: a cautionary tale about measurement error in treatment effects models. *Adv. Econ.* 27A, 1–39
30. Ferraro, P.J. and Hanauer, M.M. (2014) Advances in measuring the environmental and social impacts of environmental programs. *Annu. Rev. Environ. Resour.* 39, 495–517
31. Kundel, D. *et al.* (2018) Design and manual to construct rainout-shelters for climate change experiments in agroecosystems. *Front. Environ. Sci.* 6, 14
32. Ferraro, P.J. and Simorangkir, R. (2020) Conditional cash transfers to alleviate poverty also reduced deforestation in Indonesia. *Sci. Adv.* 6, eaaz1298
33. Tamer, E. (2010) Partial identification in econometrics. *Annu. Rev. Econom.* 2, 167–195
34. Manski, C.F. and Pepper, J.V. (2011) *Deterrence and the death penalty: partial identification analysis using repeated cross sections. NBER Working Paper 17455*, National Bureau of Economic Research
35. Rubin, D.B. (1980) Randomization analysis of experimental data: the Fisher randomization test. *J. Am. Stat. Assoc.* 75, 575–582
36. Koenig, W.D. (1999) Spatial autocorrelation of ecological phenomena. *Trends Ecol. Evol.* 14, 22–26
37. Lichstein, J.W. *et al.* (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecol. Monogr.* 72, 445–463
38. Hui, C. *et al.* (2006) A spatially explicit approach to estimating species occupancy and spatial correlation. *J. Anim. Ecol.* 75, 140–147
39. Koerner, S.E. *et al.* (2014) Plant community response to loss of large herbivores differs between North American and South African savanna grasslands. *Ecology* 95, 808–816
40. Baird, S. *et al.* (2015) *Designing experiments to measure spillover effects, second version. PIER Working Paper. No. 15-021*, Elsevier
41. Tchetgen, E.J.T. and Vanderweele, T.J. (2012) On causal inference in the presence of interference. *Stat. Methods Med. Res.* 21, 55–75
42. Rosenbaum, P.R. (2007) Interference between units in randomized experiments. *J. Am. Stat. Assoc.* 102, 191–200
43. Ho, K. and Rosen, A.M. (2015) *Partial identification in applied research: benefits and challenges. NBER Working Paper 21641*, National Bureau of Economic Research
44. Ferraro, P.J. and Agrawal, A. (2021) Synthesizing evidence in sustainability science through harmonized experiments: community monitoring in common pool resources. *Proc. Natl Acad. Sci. U. S. A.* 118, e2106489118
45. VanderWeele, T.J. and Hernan, M.A. (2013) Causal inference under multiple versions of treatment. *J. Causal Infer.* 1, 1–20
46. Kimmel, K. *et al.* (2020) Diversity-dependent soil acidification under nitrogen enrichment constrains biomass productivity. *Glob. Chang. Biol.* 26, 6594–6603
47. Isbell, F.I. *et al.* (2013) Nutrient enrichment, biodiversity loss, and consequent declines in ecosystem productivity. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11911–11916
48. Ellenberg, J.H. (1996) Intent-to-treat analysis versus as-treated analysis. *Drug Inf. J.* 30, 535–544
49. Peugh, J.L. *et al.* (2017) Beyond intent to treat (ITT): a complier average causal effect (CACE) estimation primer. *J. Sch. Psychol.* 60, 7–24
50. Angrist, J.D. and Imbens, G.W. (1991) *Identification and estimation of local average treatment effects. NBER Working Paper Series 118*, National Bureau of Economic Research
51. Kendall, B.E. (2015) A statistical symphony: instrumental variables reveal causality and control measurement error. In *Ecological statistics: contemporary theory and application* (Fox, G.A. *et al.*, eds), pp. 149–167, Oxford University Press
52. Eggers, A. *et al.* (2021) Placebo Tests for Causal Inference. Published online 2021. https://www.semanticscholar.org/paper/Placebo-Tests-for-Causal-Inference-Eggers-Tu%C3%B1%C3%B3n/c4f3e54a0908fc1efa89d149c606fac150ed5c50
53. Baird, S. *et al.* (2018) Optimal design of experiments in the presence of interference. *Rev. Econ. Stat.* 100, 844–860
54. Marion, G.M. *et al.* (1997) Open-top designs for manipulating field temperature in high-latitude ecosystems. *Glob. Chang. Biol.* 3, 20–32
55. Clark, C.M. and Tilman, D. (2008) Loss of plant species after chronic low-level nitrogen deposition to prairie grasslands. *Nature* 451, 712–715
56. Harpole, W.S. *et al.* (2016) Addition of multiple limiting resources reduces grassland diversity. *Nature* 537, 93–96

57. Wood, S.A. *et al.* (2010) Organismal traits are more important than environment for species interactions in the intertidal zone. *Ecol. Lett.* 13, 1160–1171
58. Paine, R.T. (1966) Food web complexity and species diversity. *Am. Nat.* 100, 65–75
59. Benedetti-Cecchi, L. and Cinelli, F. (1997) Confounding in field experiments: direct and indirect effects of artifacts due to the manipulation of limpets and macroalgae. *J. Exp. Mar. Biol. Ecol.* 209, 171–184
60. Smith, M.D. and Knapp, A.K. (2003) Dominant species maintain ecosystem function with non-random species loss. *Ecol. Lett.* 6, 509–517
61. Lilley, S.A. and Schiel, D.R. (2006) Community effects following the deletion of a habitat-forming alga from rocky marine shores. *Oecologia* 148, 672–681
62. Hacker, S.D. *et al.* (2019) Regional processes are stronger determinants of rocky intertidal community dynamics than local biotic interactions. *Ecology* 100, e02763