Received: 25 June 2024 | Revised: 22 April 2025 | Accepted: 6 May 2025

DOI: 10.1111/csp2.70066

CONTRIBUTED PAPER

Conservation Science and Practice
A journal of the Society for Conservation Biology
Open Access

WILEY

Selecting among counterfactual methods to evaluate conservation interventions

Tanya O'Garra[1,2] | **Rachel Martin**[3,4] | **Edwin Pynegar**[5,6] | **Claudia Polo-Urrea**[7] | **Johanna Eklund**[8]

[1]Department of Biological Sciences, National University of Singapore, Singapore

[2]Centre for Environmental Policy, Imperial College London, London, UK

[3]Conservation X Labs, Washington, DC, USA

[4]Department of Biology, University of Oxford, Oxford, UK

[5]School of Environmental and Natural Sciences, Bangor University, Bangor, UK

[6]The Biodiversity Consultancy, Cambridge, UK

[7]Independent Consultant

[8]Department of Geosciences and Geography, Faculty of Science, University of Helsinki, Finland

Correspondence
Tanya O'Garra, National University of Singapore, Department of Biological Sciences, Singapore.
Email: tanyaogarra@gmail.com

Funding information
Research Council of Finland, Grant/Award Number: 333518; Kone Foundation

Abstract

Effective conservation that benefits biodiversity and human well-being is imperative for global sustainability. Achieving this requires rigorous evaluation of conservation policies and programs to understand their causal effects on environmental and social outcomes. Counterfactual impact evaluation methods offer a robust framework for assessing intervention impacts by comparing observed outcomes with hypothetical alternative scenarios (the 'counterfactual'). Despite recent advances, a significant research-implementation gap persists in applying these methods within conservation practice. This paper is intended to help conservation practitioners, scientists, and funders respond to the growing demand for causal evaluations by providing an introductory overview of key counterfactual evaluation methods. It introduces a decision framework to guide the selection of appropriate evaluation methods according to project-specific parameters, such as project goals and timing. Application of the framework is illustrated through examples including community-managed fisheries, camera traps, and payments for ecosystem services. These examples highlight that the most appropriate evaluation method depends on various factors, such as whether the intervention can be randomized, available sample size, and data availability from both intervention and non-intervention sites. By providing a structured approach to selecting counterfactual methods for specific conservation projects, this paper aims to stimulate broader adoption of evidence-based practices in conservation.

KEYWORDS

causal inference, conservation interventions, counterfactual methods, impact evaluation

1 | INTRODUCTION

Implementation of effective conservation interventions that benefit biodiversity and people is critical for global sustainability. To achieve this, conservation policies and programmes must be informed by evidence regarding their expected impact on environmental and social outcomes. This requires an understanding of the causal effects of these interventions, that is, whether they genuinely lead to the desired effects. Counterfactual impact

evaluation methods—in which the causal effects of an intervention are identified by comparing observed outcomes with estimates of what *would* have happened in the absence of the intervention, or if implemented differently (i.e., the "counterfactual")—can provide this understanding.

Although counterfactual impact evaluation is well established as a foundation of evidence-based policy in fields such as medicine, development, and education, it has only recently started to gain traction in conservation. A recent meta-analysis of 186 counterfactual evaluations of conservation interventions (Langhammer et al., 2024) attests to the fact that counterfactual methods in conservation have increased noticeably ever since Ferraro and Pattanayak (2006) noted the scarcity of such studies. Yet a significant research–implementation gap remains. While some organizations are committed to robust impact evaluation of their projects (e.g. Craigie et al., 2015; Global Environmental Facility, 2016; McKinnon et al., 2015; Puri et al., 2020), many funding and implementing bodies largely overlook the potential for learning through impact evaluation (Craigie et al., 2015). This point has been made previously (Ferraro, 2009; Ferraro & Pattanayak, 2006; Miteva et al., 2012), with complementary studies highlighting reasons for this slow uptake, including: limited budgets; technical complexity and insufficient in-house expertise; perceptions that traditional methods based on case studies or field experience are of equal quality (Adams et al., 2019; Baylis et al., 2016; Curzon & Kontoleon, 2016). Additionally, funders and conservation organizations may be reluctant to engage in processes that might challenge expectations about project success (e.g. Asquith, 2020; Catalano et al., 2018, 2019).

However, recent events have underscored the importance of counterfactual impact evaluation methods in conservation. Chief among these was the upheaval in the voluntary carbon market due to controversies around baselines and counterfactuals used to quantify carbon credits from forest protection (Greenfield, 2023; West et al., 2020; West et al., 2023). The debates around the performance of biodiversity credits and, more recently, nature credits have also highlighted the need for counterfactual methods to ensure measurable and genuine conservation benefits (Bull et al., 2021). Relatedly, financial institutions worldwide are recognizing biodiversity's critical role in economic production (World Economic Forum, 2020), and there is a corresponding surge in business initiatives to mitigate biodiversity loss, evidenced by the growth of coalitions like Business for Nature and the creation of guidelines such as the Taskforce on Nature-related Financial Disclosures (2023). Multilateral funds for conservation such as the Global Environment Facility and the Green Climate Fund have also adopted counterfactual methods to inform the performance of their investments (GEF IEO, 2016).

In light of these push factors, and the growing interest among conservation organizations in counterfactual evaluation (Mahajan et al., 2023; McKinnon et al., 2015), this paper aims to assist conservation practitioners, scientists, and funders in making *informed decisions* about how to proceed with counterfactual impact evaluation for specific conservation projects. It is intended for those who are familiar with the concept of counterfactual impact evaluation (i.e., comparing outcomes in a "treated" group with outcomes of a similar but "untreated" group) but who are less familiar with the range of methods specifically applicable to conservation. It is particularly useful for (i) practitioners who have determined the necessity of evaluating the impact of a conservation intervention and are considering how to proceed, (ii) conservation scientists and students who are keen to use impact evaluation in their research. Acknowledging the difficulty of navigating the extensive and diverse literature on this subject, and the challenge of determining the subsequent steps, our goal is to facilitate this process by:

- Providing an introductory overview of the main counterfactual impact evaluation methods of relevance to conservation and their underlying assumptions (Section 2).
- Presenting a decision framework to help practitioners identify which method(s) might be most appropriate for the intervention being evaluated (Section 3).
- Illustrating through hypothetical and real examples how the framework can be used to decide which methods are appropriate, given different constraints and their assumptions (Section 4).

We note that Section 2 is relatively technical and assumes some familiarity with statistical concepts. Readers who are already acquainted with the foundational aspects of counterfactual methods may choose to skim or skip the more detailed discussion and head straight to the Decision Framework (Section 3). However, we strongly encourage all readers to review Section 2.1, which outlines essential concepts such as treatment assignment, confounding, and strategies for identifying confounders.

A central tenet underlying the framework presented here is that there is no 'best method' for evaluating impact. Rather, there are methods that are appropriate for the task at hand. Appropriateness may depend on many factors, such as data availability, study timescale, and available sample size. For example, methods that are appropriate to evaluate large-scale, costly conservation projects may not be appropriate for small pilot projects, and vice versa.

We highlight that this paper does not provide guidance about how to implement different methods. Each method entails different procedures, which have been covered in depth in existing resources that are cited throughout this paper (see especially Section 2.1). Additionally, the paper does not provide guidance about the preliminary steps underpinning impact evaluation, such as formulating a theory of change, determining the need for implementing impact evaluation, or identifying evaluation questions. These steps are covered in a companion piece in this Special Issue ('Introduction to Impact Evaluation in Conservation' by Neugarten et al., 2025), to which interested readers can refer.

Our intention with this paper is to provide a practical framework to help readers assess the feasibility of different impact evaluation approaches, whether for planning new evaluations, improving existing studies, or engaging with the literature. It offers guidance on making key decisions, such as whether to seek expertise, allocate resources, or gather additional data. The framework also highlights how to incorporate evaluation considerations early in project planning, particularly around data collection and program design. We expect that, over time, counterfactual methods will be increasingly used strategically to learn from past experiences, replicate successes, correct failures, and avoid mistakes in project design and implementation.

2 | OVERVIEW OF COUNTERFACTUAL METHODS

2.1 | Basic principles of counterfactual impact evaluation

Counterfactual methods include a wide range of approaches that are used to explore cause-effect relationships between interventions and outcomes. They do this by comparing what actually happened after an intervention to estimates of what would have happened in the absence of the intervention or a different version of it (Rubin, 1978, 2005). These "what-if" scenarios are known as *counterfactuals*.

Given that we can never simultaneously observe alternative states of the world (with and without the intervention for any given unit receiving the intervention), but only observe what actually occurred, researchers rely on assumptions to identify plausible counterfactuals. A common approach is to use sites or units of study where the intervention did not take place (or did so differently). However, if these comparison groups are not selected randomly (see Section 2.1)—or without a clear and intentional sampling strategy—then they might not provide valid counterfactuals. This is because other factors, unrelated to the intervention, could influence the outcome and introduce *confounding*.

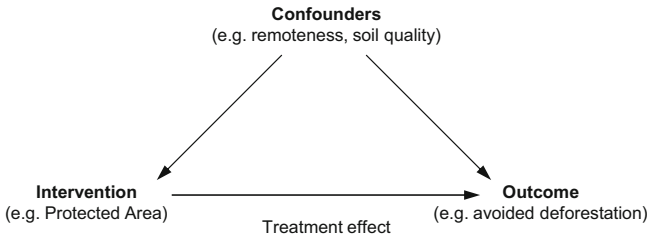

FIGURE 1 Simple causal diagram. Arrows indicate direct causal effects between variables. Here the causal (treatment) effect of an intervention (e.g., establishing a Protected Area) on the outcome (e.g., avoided deforestation) is confounded by common causes (e.g., remoteness). For example, protected areas (PAs) are often established in remote locations, while deforestation is more common near roads and urban areas (Joppa & Pfaff, 2009). Without controlling for remoteness as a confounding variable, it could look like PAs are more effective in preventing deforestation than is actually the case. Confounders should always predate the intervention and the outcome(s), and not be influenced by either of them. Confounding variables can include socioeconomic status, geography, or other environmental variables that influence both the implementation of the intervention and the outcomes being measured.

Confounding describes a situation in which both the intervention and the outcome of interest have a common cause (see Figure 1). These common causes ('confounders') are factors that influence both the assignment, or adoption, of the intervention and the outcome(s) of interest. Unless properly identified and accounted for, confounders can distort the true relationship between the intervention and the outcome(s), leading to incorrect conclusions about causality.

All counterfactual methods depend on the critical assumption: that the relationship between the intervention and outcome of interest is not affected by confounders (Morgan & Winship, 2015). To satisfy this assumption, potential confounders must be identified and controlled for. This can be achieved through randomization or statistical adjustments. In qualitative studies, case selection based on rigorous implementation of a "most similar" approach which accounts for confounders can be applied (Plümper et al., 2019). By controlling for confounders in any of these ways, evaluators can more confidently attribute observed changes in outcomes to the intervention itself, rather than to the influence of confounders. Therefore, careful identification and adjustment for confounding variables are crucial steps in conducting rigorous impact evaluations of conservation interventions (Dee et al., 2023; Ferraro & Hanauer, 2014).

2.1.1 | How to identify confounders?

To identify confounders, evaluators must carefully determine how an intervention is (or will be) assigned. This means identifying selection criteria, decision processes, and/or motivations that together explain how a treatment is (or was) assigned to different 'units' (i.e., plots of land, households, individuals, etc.). These influences are collectively known as the "treatment assignment mechanism".

If an intervention is not randomly assigned, then evaluators must engage in identifying this treatment assignment mechanism. In practice, this means gathering detailed information about how an intervention was assigned and the sources of information and/or data used to guide those decisions, often through interviews with key personnel.

For interventions involving voluntary participation (self-selection), identifying confounders is more challenging, as it requires determining the individual or community factors that influence participation. In these cases, domain expertise and a comprehensive review of relevant literature—essential for all studies—are particularly critical for ensuring valid evaluation. For all counterfactual studies, there should be explicit and detailed reporting of the processes through which confounders were identified by evaluators, how they are expected to affect treatment assignment and outcomes, and what data is used to control for them.

2.2 | Broad impact evaluation methods

There are three broad categories of methods available for counterfactual evaluation of impact, which are outlined below:

- **Experiments** involve the random assignment of an intervention—or versions of an intervention—over eligible units (e.g., sites, villages, and households). Some units randomly receive the intervention (the "treatment" group) while others do not, or receive a different version of it (the "control" group). If designed and implemented correctly, randomization greatly reduces or eliminates the likelihood of systematic differences between treatment and control groups (other than the intervention). This ensures that observed changes are attributable to the intervention rather than to confounders (Ferraro, 2012; Gertler et al., 2016; Glennerster & Takavarasha, 2013). In conservation, experiments include field-based experiments (e.g., Jayachandran et al., 2017; Pynegar et al., 2018, 2021), as well as lab and online experiments typically used in social marketing studies (e.g. Blake et al., 2023; Dunn et al., 2020). Experiments are always designed before the intervention is implemented.

- **Quasi-experiments** use statistical approaches to eliminate confounders and other rival explanations for the observed treatment effects. They are typically used when randomization is not possible, which is usually the case for interventions that are already in place and which were assigned or adopted non-randomly. There are a range of different quasi-experimental designs, such as difference-in-differences, statistical matching, and synthetic control methods. The main differences between these methods are outlined in Table 1. See Ferraro and Hanauer (2014) and Greenstone and Gayer (2009) for detailed insights into quasi-experimental methods and how they compare to experimental approaches.

- **Qualitative attribution methods** offer insights into how interventions lead to potential impacts by examining mechanisms, contextual factors, and participant experiences (Zavaleta Cheek et al., 2023). These methods encompass various approaches, all of which gather empirical evidence and validate theories of change against alternative explanations using techniques such as interviews, case studies, and document analysis. Qualitative attribution methods are useful for understanding contextual factors that quantitative measures may overlook. While they may not provide statistical effect sizes, qualitative attribution methods such as process tracing, general elimination methodology, contribution analysis, and realist evaluation have been utilized to assess the impact of conservation and environmental interventions (Gregory-Smith et al., 2017; Rana & Miller, 2021; Sabin et al., 2019; Salazar et al., 2018). Qualitative attribution methods vary in the extent they can be considered counterfactual, with some specifically designed to eliminate alternative explanations and others primarily designed to systematically trace emerging patterns that could be linked to causation through different mechanisms (Zavaleta Cheek et al., 2023). The reliability of qualitative methods for making valid inferences is strongly linked to the selection of comparable cases that help control for confounding factors (Plumper et al. Plümper et al., 2019).

Overall, these broad categories are mainly distinguished by how the intervention is implemented (randomization vs. all other), and whether they produce quantitative versus qualitative results. Quantitative methods are well-suited to isolating specific intervention effects, while qualitative methods tend to be used to evaluate mechanisms when direct (quantitative) observations of mechanisms are lacking. However, quantitative methods can

also be used to quantify mechanisms (e.g., Ferraro & Hanauer, 2015; O'Garra et al., 2023).

All these categories comprise a range of different approaches. In this paper, we provide more detail about different quasi-experimental approaches (Section 2.3), because these are well-suited to evaluate conservation interventions and are likely to see wider application in this field. Quasi-experiments can, under certain conditions (see Sections 2.3 and 3.2), enable retrospective evaluation of existing interventions and prospective evaluation of interventions in which randomization is not feasible. Their flexibility makes them particularly valuable in conservation, where randomization is often impractical (see Section 3.2, Step 2B).

For qualitative attribution methods, a recent paper by Zavaleta Cheek et al. (2023) provides a comprehensive overview of the different approaches, with illustrative examples; hence, we do not repeat this information here and instead guide readers to that resource. For detailed insights into variations in experimental designs, see Duflo et al. (2007) and Glennerster and Takavarasha (2013).

We acknowledge that there may be other ways of classifying methods; however, a review of classifications used across different fields (Appendix S1) suggests our classification aligns well with those reported in the literature.

2.3 | Quasi-experimental methods

Table 1 summarizes key quasi-experimental methods for conservation, focusing on commonly used methods in conservation supported by an extensive body of knowledge (e.g., journal papers, reports, and presentations) that new evaluators can draw on. These include: difference-in-differences (DID), matching, and synthetic control method (SCM). We also include two promising methods that are rarely used in conservation: Interrupted Time Series (ITS), which is useful for analyzing data from treated sites when there is no comparable data from non-treated sites, and Regression Discontinuity Design (RDD), applicable for interventions that are implemented based on a threshold value of a continuous variable (see Table 1 for examples and definitions of each method).

Table 1 shows that each method varies mainly in terms of how the counterfactual is constructed, the underlying assumptions ("when to use"), and the types of data used. In terms of data requirements, DID (a more rigorous version of a before-after-control-impact design, that accounts for confounders in selecting the controls), SCM, and ITS all use longitudinal data (involving repeated observations of the same variables from the same sample over time), although the length of the required time-series varies by method; DID only requires pre-treatment observations to validate the 'parallel trends' assumption (see Table 1 for definition), and otherwise only needs a single pre-treatment (baseline) and single post-treatment outcome observation to estimate impact. SCM and ITS however need long pre- and post-treatment time-series, although SCM requires multiple non-treated units for creation of a comparison group, whereas ITS can be used without any non-treated units as long as the time-series is sufficiently long to establish robust evidence of trends.

Each method also differs in how the comparison group(s) (the counterfactual) is generated and the assumptions underpinning this process. Matching selects comparable treated and untreated units based on observable confounders (i.e., confounders that can be measured and controlled for). The matched untreated units represent the counterfactual. DID involves comparing outcomes of treated and untreated units before and after an intervention, based on the assumption that trends in the outcomes of untreated units reflect what the treated units would have experienced without the intervention (i.e., the counterfactual). This is partly verified by checking parallel pre-treatment trends; DID controls for unobservable confounders (i.e., which cannot be measured or controlled for) that do not change over time. SCM controls for all unobservable confounders, including time-varying ones (Abadie et al., 2015), by using longitudinal data to create a 'synthetic control' from weighted non-treated units that match the treated units' pre-treatment trends. ITS uses long time-series data with clear trends from intervention sites only to generate counterfactuals by extrapolating from the pre-treatment trend; this assumes the pre-treatment trend would continue without an intervention. RDD generates a counterfactual by assuming that treatment assignment is effectively random for units (e.g., farm owner, household, and individual) just above and below the eligibility cutoff (e.g., farm size, income, and years of ownership); see Table 1.

Notably, each design has different assumptions for causal inference, which determine the causal effect being measured (known as the "estimand") (Ferraro & Hanauer, 2014). For example, RDD only estimates impact for units close to an eligibility cutoff, so results are not generalizable to units far from this cutoff, while an experiment implemented among the general population produces estimates of average impact that are generalizable over that population.

Although the various quasi-experimental methods are presented separately in Table 1, they can be combined in

TABLE 1 Overview of quasi-experimental methods.

	Overview	When to use	Data required	References
Difference-in-differences (DID)	Used to estimate an intervention's impact by comparing changes in outcomes over time between treatment and control groups. It calculates the difference in outcomes between the groups after the intervention, then subtracts the pre-intervention difference from this. This isolates the effect of the intervention. It assumes that, in the absence of the intervention, both groups would have followed similar ("parallel") trends. Example: evaluate the impact of a PA on deforestation by comparing forest loss trends in protected and unprotected areas before and after designation.	Use when, (i) treatment and non-treatment groups experience the same conditions during the evaluation period; (ii) the effects of unobserved confounders (those you do not have data on) are constant over time; (iii) outcomes for treatment and non-treatment groups would have followed similar trends if there had been no treatment (needs longitudinal pre-intervention data on outcomes from both groups prior to identify parallel trends visually).	Data on outcomes before and after treatment on both treated and non-treated units. To check if the groups had similar trends before the treatment (see Assumptions), it's helpful to have data from multiple time periods before the intervention.	Method: Gertler et al. (2016), Fredriksson and de Magalhães Oliveira (2019) Applications: Wauchope et al. (2022); Francini-Filho and Moura (2008); Feng et al. (2021)
Matching	Used to create two groups (treatment and non-treatment) that are similar with regards to confounders. This ensures fair comparisons by controlling for the effects of confounders. It can be used to generate samples for data collection (see Stuart & Lalongo, 2010) or be applied to existing data for analysis after an intervention. Example: evaluate the impact of a PA on deforestation by comparing protected and unprotected sites that have been selected using matching so that they are as similar as possible with regards to confounders.	Use when (i) all confounders are observed (i.e., you have pre-treatment data on them) and controlled for (or if their effects are captured by observed confounders that are controlled for); (ii) there are enough treatment and non-treatment units to select your comparison groups from. Matching does not require longitudinal data, but only data from a single point in time shortly before treatment.	Pre-treatment data on confounders from treated and non-treated units is needed to select groups for comparison. For analysis, combine with DID. If only post-treatment data available, use single differences (e.g. O'Garra et al., 2023) but extra care is needed to identify and control for all confounders.	Method: Schleicher et al. (2020); Stuart (2010); Stuart and Lalongo (2010) Applications: Devenish et al. (2022); Andam et al. (2008); O'Garra et al. (2023)
Synthetic control method (SCM)	Used to create a comparison for a single or a few treated units. It combines non-treated sites into a realistic comparison unit (the "synthetic control"). Interventions must occur at a specific point in time, although staggered timings can work if clearly defined and not spread out over a long period. Useful to evaluate interventions affecting a single or few treatment units. Example: evaluate if a single REDD+ project caused additional reductions in deforestation compared to a synthetic control.	Use when, (i) you have longitudinal data (pre- and post-intervention) from enough untreated sites to create a realistic comparison (synthetic control); (ii) outcomes for treated units can be predicted by a simple average of untreated units ("linearity" assumption), (iii) there are no major events or changes after the intervention that affect treated and comparison groups differently	Data is needed on outcomes and factors that are expected to influence outcomes for treated unit(s) and a set of non-treated comparison units for many pre-intervention periods. Post-intervention data over multiple time periods is essential.	Method: Kreif et al. (2016); Abadie et al. (2010) Applications: West et al. (2020); Jones et al. (2020); Sills et al. (2015); Sharma et al. (2023)
Regression discontinuity design (RDD)	Used to estimate the impact of an intervention by comparing outcomes for units just above and just below a cutoff point, which is used to decide who or what is eligible to receive the intervention. RDD only estimates impact for units close to the cutoff (a "local" average treatment effect), not the entire population. Example: farm size is used to determine eligibility for Payments for Ecosystem Services; compare farms just above and below the farm size cutoff.	Use when, (i) there is an "eligibility" variable that is used to rank units on a continuous scale; (ii) there is a specific, clearly defined threshold (cutoff) on this scale and all units are assigned above or below it; (iii) subjects (units) cannot manipulate the eligibility variable (they cannot change their score to qualify), other-wise the cutoff cannot be treated as a random assignment device.	Data is needed on the eligibility variable, which must be continuously distributed. Outcome data is needed for units on either side of the cutoff. Data on factors that might influence outcomes can help control for additional effects.	Method: Gertler et al. (2016); Wuepper and Finger (2023) Applications: Alix-Garcia et al. (2018); Baragwanath and Bayi (2020)

TABLE 1 (Continued)

	Overview	When to use	Data required	References
Interrupted time series (ITS)	Used to estimate impact of an intervention when there is long-term data, pre- and post-intervention. The counterfactual is constructed by extending the pre-intervention trend, and comparing this to the observed trend in outcomes. Especially useful when there is only data from treated groups. Example: to evaluate if a hunting ban reversed a downward trend in an elephant population (looking at change in trend, rather than average count)	Use when, (i) there is a trend in the outcome variable over time; (ii) the pre-intervention trend remains constant in the absence of an intervention; (iii) the intervention occurs at a specific moment in time, (iv) intervention effects occur soon after exposure; Additionally: analyses must account for auto-correlation (including seasonal effects) (see Turner et al., 2021)	Pre- and post-longitudinal data on outcomes and factors that are expected to influence outcomes, ideally for as long a time-series as possible.	Method: Wauchope et al. (2021); Turner et al. (2021); Schaffer et al. (2021) Applications: Ota et al. (2023); Fox et al. (2022)

many cases to enhance the validity and robustness of causal inference by addressing the limitations inherent in individual approaches. For example, DID is often combined with matching (e.g., Andam et al., 2008; Devenish et al., 2022) and has also been used in combination with SCM (e.g., Arkhangelsky et al., 2021). Matching can be combined with most of the other approaches (e.g., Kellogg et al., 2019; Linden, 2018).

3 | INFORMED OPTIONS DECISION FRAMEWORK

3.1 | Scope

Before using this framework to select an appropriate impact evaluation method, users should have already completed essential preliminary steps, such as ascertaining the need for an evaluation, formulating a theory of change, clearly specifying the evaluation questions, and identifying outcomes and associated indicators (see Neugarten et al., 2025). Although it is essential to address them early on, these preliminary steps may be revisited as data availability and analytical methods are identified.

We also reiterate that, before embarking on any counterfactual evaluation, the evaluator must carefully identify potential confounders (see Section 2.1) and obtain baseline data on these confounders to control for their effects.[1]

In establishing the starting point and focus of our framework, we recognize that the primary concern for our audience is likely to be whether their project makes a difference. Additionally, budding evaluators might be wondering how to best utilize the data they have already collected to evaluate project efficacy. These two main considerations set the scope for the framework presented here, which is on methods that allow the user to assess the impact of an intervention, while considering data availability.

We acknowledge that evaluators may be interested in addressing other questions such as: how does the intervention impact different groups? What factors enhance/diminish the impact of the intervention? What is the effect size of different mechanisms[2] on the outcomes? Our proposed framework does not consider these additional questions, and interested readers can instead refer to Ferraro et al. (2011) and Ferraro and Pressey (2015) for studies of moderators and heterogeneity; O'Garra et al. (2023), Wiik et al. (2020), Reimer and Haynie (2018), Ferraro and Hanauer (2015) for studies quantifying the hypothesized mechanisms to impact; and for a technical overview of heterogeneity in impact evaluation, see Vivalt (2015).

3.2 | Decision framework

The informed options decision framework is divided into multiple sections, outlined in Figure 2. The evaluator can move through the framework sequentially; however, an iterative approach will be more appropriate and useful in most cases. The decision framework is organized around broad questions that evaluators will need to consider—relating to the study aims (broadly), the intervention, data availability, sample sizes, and technical complexity and effort. Details about each of these steps, and how to move through them, are provided below (Steps 1–4).

3.2.1 | Step 1: broad methods

The decision process for determining which broad method(s) might be suitable to evaluate a project is given in Decision Matrix 1 (Table 2).

The first question asks whether the intervention has already been implemented. Answering this, the user moves to question 2, which asks whether randomization, or a randomly staggered roll-out, is feasible. Next, the evaluator must consider the study aims (question 3). The need for a quantitative (statistical) estimate of effect size may depend on the overarching goal of the study. For example, if the goal is to assess whether to scale a small pilot, a broad assessment of the direction of effect may suffice. Indeed, small pilot studies typically lack statistical power to estimate effect sizes (see Section 3.2, Step 3). If the goal, however, is to identify the contribution of an initiative to quantitative targets, then statistical effect sizes will be needed.

Notably, every cell in Decision Matrix 1 includes quasi-experiments as an option. This is because many quasi-experimental approaches have been developed to accommodate different assumptions about how an intervention is assigned and a wide range of data settings. The decision process for identifying suitable quasi-experimental methods is addressed in Step 2A of the Decision Framework (Figure 2) using Decision Matrix 2 (Table 3). For evaluators who have selected experimental methods as an option, key questions to help move forward are identified in Step 2B. For information to guide the choice between different qualitative methods, see Zavaleta Cheek et al. (2023), which outlines the process for selecting between various qualitative attribution methods.

3.2.2 | Step 2A: quasi-experimental methods

Decision Matrix 2 (Table 3) guides selection of suitable quasi-experimental methods by focusing on key questions about data availability for treatment and control units. Before proceeding, however, evaluators must ensure that they have carefully identified all confounders and collected suitable data to control for their effects (see Section 2.1). This crucial step ensures that estimated impacts can be causally attributed to the intervention, not confounding factors. Additionally, evaluators must identify the unit(s) of observation for their analysis. While this is typically addressed during the preliminary steps (outlined in Section 3.1), it is worth revisiting now. Determining which units of observation can be measured is important, as outcome data should be collected for units that allow for valid conclusions relating to the goals of the study. For example, if the evaluator is interested in the effect of a village-level intervention on livelihoods, then outcomes measured at the national level would not be useful. Thus, when using Decision Matrix 2, evaluators must consider not only what data is available, but at what scale or level, ensuring it aligns with the study goals.

As shown, the available data determines which quasi-experimental method can be used, if any. Long time-series data covering both pre- and post-treatment time periods for treated and non-treated units allow for the widest range of methods. The key constraint in choosing between methods is the availability of data for non-treated sites. If long pre- and post-intervention time-series data is only available from treated sites, then ITS is the only possible method. If the evaluator only has limited time-series data or only single observations from treated sites, they are advised to revisit Step 1 (Figure 2) and use Decision Matrix 1 (Table 2) to explore possible alternative methods.

If only post-intervention outcome data is available, then the evaluator could use RDD for interventions involving assignment using cutoffs, or they might be able to use matching. In both cases, any assessment of impact will be cross-sectional ("single differences").

If there are multiple outcomes of interest, and/or different data on outcomes, then the evaluator must *consider one outcome (and/or dataset) at a time*, and work through this matrix several times. This may result in different options for evaluating impact for different outcomes.

Having worked through Decision Matrix 2, the evaluator may now face a variety of choices. If no quasi-experimental options are possible, they are advised to revisit Step 1 of the Decision Framework (Figure 2) and, using Decision Matrix 1, identify whether and how they can adjust their choices. On the other hand, if the evaluator has identified one or more possible quasi-experimental methods that they can use, they should proceed to Step 3 to assess the sample size needed for their study.

3.2.3 | Step 2B: experimental feasibility

If the evaluator has identified experiments as a possible option in Step 1 (Figure 2) using Decision Matrix 1 (Table 2), they must consider several factors to determine whether an experiment is feasible:

1. Unit of randomization—this describes the entity allocated randomly to treatment or control in an experiment. It must be selected by the experimenter and may differ from the unit to which the treatment or intervention is applied, resulting in a cluster RCT (e.g., Grillos et al., 2019). For instance, if an intervention targets households, but randomization occurs at the village level, all households within a village will either receive the intervention or not, based on the random assignment of villages to treatment or control groups.

Experimenters must consider the following when determining a randomization unit:

- Sample size and expected effect size: randomizing interventions over smaller units (e.g., households) will often provide a larger potential sample size compared to larger units (e.g., villages). Larger sample sizes enhance the power to statistically detect true effects; conversely, larger (expected) effect sizes require smaller samples to detect true effects. Hence, the randomization unit should be chosen to secure a sufficient sample size to estimate an effect. Considerations about sample size and expected effect sizes are critical for all quantitative counterfactual methods; we cover this in detail in Step 3.
- Spillover effects: these occur when the impact of an intervention extends beyond the group targeted by the intervention, violating the Stable Unit Treatment Value Assumption—which holds that outcomes for

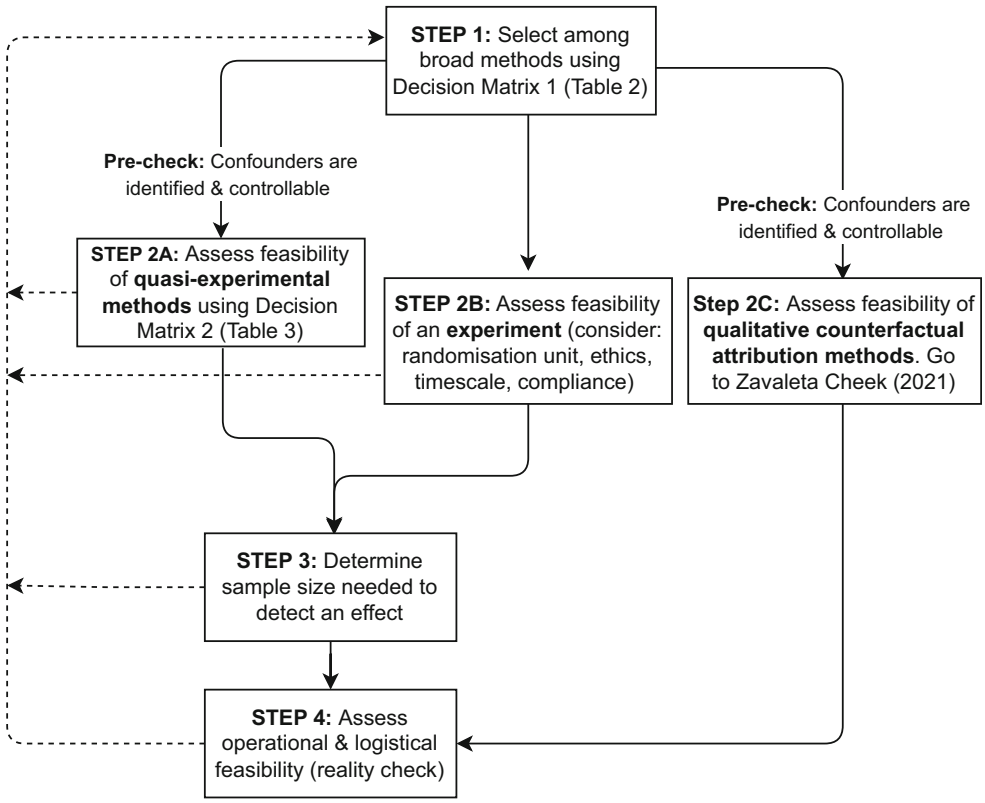

FIGURE 2 Flow diagram showing informed options decisions framework. Step 1 helps evaluators understand whether they can implement experimental, quasi-experimental, and/or qualitative methods for their study with help from Decision Matrix 1 (Table 2). Depending on the broad method(s) selected in Step 1, evaluators can then move to one *or multiple* of Steps 2A, 2B, and/or 2C. Step 2A presents Decision Matrix 2 (Table 3) which has key information needed to assist with the selection of appropriate quasi-experimental methods (if any); Step 2B guides evaluators in identifying the feasibility of experimental methods. For Step 2C, evaluators must use Zavaleta Cheek et al. (2023) to assess the feasibility of qualitative methods. Sampling and sample size considerations are considered in Step 3, followed by advice about how to proceed given different technical demands of the various methods, costs, and other considerations (Step 4). Dashed arrows from each step indicate that the reader must return to Step 1 if no feasible method is found at that step.

TABLE 2 Decision Matrix 1 for choosing among broad approaches (Step 1 in Figure 2).

		Decision Matrix 1		
		1. Has the intervention already been implemented?		
		NO		YES
	2. Can you randomize or randomly stagger roll-out?	YES	NO	n/a
3. Study aims to evaluate…	..effect of an intervention (statistical effect size not needed)	All methods possible	Quasi-experimental Qualitative	Quasi-experimental Qualitative
	..effect of an intervention AND statistical effect size	Experimental Quasi-experimental	Quasi-experimental	Quasi-experimental

TABLE 3 Decision Matrix 2 for choosing among quasi-experimental methods (Step 2A in Figure 2).

		Decision Matrix 2		
		1. What outcome data do you have and/or can you get?		
		Pre- & post-intervention time series	Pre- & post-intervention data (single observation or time-series)	Post-intervention data only (single observation or time-series)
2. You have this data for…	Treated & non-treated sites	All methods possible	Matching Difference-in-differences[1] Regression discontinuity[2]	Matching (single difference analysis) Regression discontinuity[2]
	Treated sites only	Interrupted time series	No methods possible *(revisit Step 1: Decision Matrix 1)*	No methods possible *(revisit Step 1: Decision Matrix 1)*

[1]DID typically requires multiple pre-intervention observations to validate the 'parallel trends' assumption. However, there are ways of making this assumption more credible, including combining DID with matching (Mackenzie 2023).
[2]Can only use RDD if treatment is assigned according to a threshold value for an eligibility variable (see Table 1).

any unit depend only on whether it received treatment, and not on the treatment status of other units (Kimmel et al., 2021; Rubin, 1980). This can lead to biased results, so it is important to set up experiments that minimize spillover effects. Randomization strategies should thus avoid including intervention units that are close to each other, or that interact in any way, as these are more likely to suffer from spillover. For more details, see Ferraro et al. (2019) or Pynegar et al. (2025). Spillover effects can affect all counterfactual methods; hence, they are discussed in more detail in Step 3.

- Evaluation-driven effects: awareness of being in an experiment can alter participants' behavior; this can affect the treatment group (*Hawthorne effect*) and/or the control group (*John Henry effect*) (Duflo et al., 2008; Peters et al., 2018). For instance, in a randomized payment-based scheme within a community, individuals in the control group might change behavior due to resentment, while those receiving treatment might respond positively simply because they were chosen for treatment. Evaluation-driven effects are largely unstudied in conservation, except for Martin et al. (2014) who found that 44% of control respondents reported changing their behavior in anticipation of future program participation. However, such effects have been frequently found in development and agricultural studies (e.g., Bulte et al., 2014; Treurniet, 2023; Zwane et al., 2011). If it is not possible to randomize an intervention without alerting participants to their involvement, for ethical or other reasons, then transparent randomization procedures can reduce bias associated with these effects (Aldashev et al., 2017).

2. Ethics: are there unresolvable ethical issues associated with potential randomization of the intervention? The ethics of randomizing access to a program have been much discussed, with issues including risks of participants being instrumentalized by evaluators, ability of all participants to provide informed consent, equipoise (that evaluators should be genuinely unsure as to which of the treatment arms is likely better), and cultural inappropriateness of randomization in some contexts (Fives et al., 2013). For example, randomizing payment-type

interventions among households within a community could lead to internal conflict if control-group households are unhappy that their allocation to the control group means that they do not receive the payment in question. If there is no way of resolving such issues to the satisfaction of potential participants, with them granting free, prior, and informed consent, then the evaluator should not proceed with the experimental approach.

3. Timescale of intervention: many processes of change are slow, and the amount of time that passes between the intervention implementation and the expected endline data collection must be realistic to detect changes. Evaluators therefore must assess whether they can run the experiment over the required time needed to detect a change. Relatedly, denying the control group the intervention for a long period of time may call into question the ethics of the experiment (see point 2).
4. Risk of noncompliance: evaluators must consider whether it is possible that experimental participants will not participate as intended in the experiment, either through not following through and dropping out of the experiment, or doing the opposite of what the intervention prescribes. There are several reasons why this may happen, including long timescales leading to frustration (see 3 above) and spillover- or evaluation-driven effects (see 1 above). Regardless, noncompliance leads to both reduced effective sample size and can reintroduce systematic differences between experimental and control groups if certain types of participants are more likely to not comply. Evaluators must therefore allow for the possibility of noncompliance when calculating sample sizes and, if it is substantial, consider whether additional analytical approaches may be needed to ensure it is not reintroducing bias (e.g., Bell et al., 2013; Molina-Millán & Macours, 2025).

The evaluator will have to carefully consider all these questions to assess the feasibility of an experimental approach. If they cannot identify a suitable randomization unit, and/or cannot proceed with an experimental approach due to ethical reasons, timeframe constraints, and/or noncompliance issues, they must consider other approaches, such as quasi-experimental designs (return to Decision Matrix 2) or qualitative designs (Zavaleta Cheek et al., 2023). If randomization is feasible, then they are advised to proceed to Step 3 to assess the minimum sample size needed for the study.

3.2.4 | Step 3: sample size determination

How many treatment and control units are available to use in your study? This is mainly determined by the population of available treatment and control units but can be constrained if spillover effects are present. These occur when the impact of an intervention extends beyond the immediate target group receiving the intervention, violating the Stable Unit Treatment Value Assumption, defined above (Kimmel et al., 2021; Rubin, 1980). Spillover effects can be biophysical (e.g., species movements from protected to unprotected areas) or socially mediated (e.g., communication between treated and untreated groups), and they can be positive or negative. For example, a Marine Protected Area may lead to increased fish populations migrating into adjacent areas (a positive spillover) but might also lead to displaced fishing effort (a negative spillover). If spillover effects are present and not accounted for, then estimates of the causal effect of an intervention will be biased. To manage spillover effects, researchers often exclude potential control units that lie within a certain distance (a buffer) of treatment sites. Spillover effects may vary in magnitude across the landscape (Robalino et al., 2017), meaning that larger buffers around treatment units are more likely to ensure that spillover effects do not bias the evaluation. However, extending the buffer reduces the pool of potential control units for the evaluation.

Having identified a suitable sampling frame, the evaluator may now ask: are there enough treatment and control units to detect a treatment effect *if it exists*? How big does the sample need to be to statistically identify a true effect? To answer these questions, a power analysis must be conducted to identify the minimum sample size needed. Underpowered studies can lead to false conclusions, such as finding an intervention has no effect, when in fact it does. Larger sample sizes improve the accuracy and power of the study but require more time, expense, and effort. Thus, identifying the minimum sample size needed to detect a true effect is crucial.

While current software simplifies power analysis, selecting the right parameters and sampling approach can be complex and require advanced technical skills, especially with clustered interventions. Clustering is common in conservation studies, where observation units are nested within treatment units (e.g., land-use pixels within Protected Areas or villages within conservation projects). We strongly advise new evaluators to consult or collaborate with experts in power analysis. However, to exemplify how a power analysis works, we present an example based on the simplest possible type of impact evaluation—a balanced experimental design (found in Appendix S2). Although this example provides an intuitive understanding of how power analysis works, most conservation studies will not be so simple, and power analysis will be more complex. For additional information on sample size calculations, we suggest Gupta et al. (2022).

3.2.5 | Step 4: reality check (operational and logistical feasibility)

If the evaluator has reached this Step of the framework, they have likely selected a suitable method to evaluate the project, identified a minimum sample size (for the quantitative approaches), and are nearly ready to implement the intervention or collect data. At this stage, they must do a 'reality check' and ask:

- *What kinds of decisions will be informed by this counterfactual evaluation*?

This question should be considered as a preliminary step, but is also crucial to confirm now. Key questions include: is there insufficient evidence about the intervention's effect? Is the intervention high risk—meaning, could failure or unintended consequences have serious impacts on people, biodiversity, or both? Can the intervention be scaled up to other contexts or locations? Would results of this evaluation be useful for other projects (i.e., are the findings generalizable)? Is the intervention large in spatial scale or budget? Finally, is evidence of additionality required, such as in the case of carbon offsets? Positive answers suggest a counterfactual impact evaluation is useful and perhaps even necessary.

- *Do we have* the *technical expertise needed, and/or can we access it*?

Many conservation organizations face budget constraints for both programmatic efforts and evaluation mechanisms (Kleiman et al., 2000). Limited in-house technical capacity may require collaboration with experts, such as university researchers, students, or specialized consultants.

- *Can we afford it*?

Impact evaluation costs vary based on program size, study design, effect size, sample size, and outcomes of interest. The cost of quasi-experimental and qualitative studies essentially corresponds to the salaries of researchers, although data collection costs can still represent a substantial portion. In contrast, randomized trials require at least two rounds of data collection in the field.

If the reality check fails, it is worth considering if the chosen approach can be tweaked or if a different approach is possible (i.e., revisit Decision Matrix 1), before ruling out rigorous impact evaluation for the current project.

4 | ILLUSTRATIVE EXAMPLE

Here we present an illustrative example of how the decision framework can be used, as applied to a hypothetical case study, with two additional examples in Appendix S3. The first of the additional examples involves a real-world application with Conservation X Labs, n.d. (https://conservationxlabs.com/), which is assessing methods to evaluate the impact of their AI-enabled camera traps on various outcomes, such as response time for invasive species detection and removal. The second focuses on a hypothetical payment-for-ecosystem-services program for forest conservation. These examples highlight how the framework can be applied in both hypothetical and real-world scenarios, demonstrating its practical utility and potential to guide effective conservation practice.

4.1 | Community-based marine management example

In this example, an impact evaluation of an existing community-based marine management initiative is being undertaken. The initiative has been operational for five years in 150 municipalities in a tropical region, and the purpose of the evaluation is to generate rigorous evidence of changes (if any) in fish biomass. The evaluator has already completed the preliminary steps that precede any impact evaluation, including generating a theory of change and identifying the evaluation question, potential outcomes, and indicators. At this stage, she is faced with the question of which methods are appropriate for this evaluation.

Starting at Step 1 of the Decision Framework (see Figure 2), the evaluator identifies that a quasi-experimental approach is the only option because: (i) the intervention has already been implemented, (ii) project funders have requested a quantitative estimate of impact on average fish biomass at the municipal level (Figure 3).

The evaluator then turns to Step 2A of the Decision Framework to identify possible quasi-experimental approach(es) to use. Before proceeding, she ensures that she has identified the treatment assignment mechanism and all potential confounders from experts, team members and the literature, and has sourced suitable baseline data for these covariates (see Section 2.1). Using Decision Matrix 2, the evaluator notes that fish biomass data has been collected from intervention sites (municipal waters) since the beginning of the intervention, but there is no time-series data prior to the intervention, only a single baseline observation. This existing data can be broadly categorized as "pre-intervention single observation & post-

Decision Matrix 1: selecting among broad approaches

		Decision Matrix 1		
		1. Has the intervention already been implemented?		
		NO		YES
	2. Can you randomize or randomly stagger roll-out?	YES	NO	n/a
3. Study aims to evaluate…	..effect of an intervention (statistical effect size not needed)	All methods possible	Quasi-experimental Qualitative	Quasi-experimental Qualitative
	..effect of an intervention AND statistical effect size	Experimental Quasi-experimental	Quasi-experimental	Quasi-experimental

Step 1: evaluator uses Decision Matrix 1 to select among broad approaches to evaluate community-based marine management project. Only quasi-experimental methods are possible given the intervention timing and project goals.

STEP 2A: using Decision Matrix 2 to select quasi-experimental methods – first pass

		Decision Matrix 2		
		1. What outcome data do you have and/or can you get?		
		Pre- & post-intervention time series	Pre- & post-intervention data (single observation or time-series)	Post-intervention data only (single observation or time-series)
2. You have this data for…	Treated & non-treated sites	All methods possible	Matching Difference-in-differences Regression discontinuity	Matching (single difference analysis) Regression discontinuity
	Treated sites only	Interrupted time series	No methods possible (*revisit Step 1: Decision Matrix 1)*	No methods possible *(revisit Step 1: Decision Matrix 1)*

Step 2A: evaluator uses Decision Matrix 2 to assess whether quasi-experimental approaches are feasible to evaluate the project's impact on fish biomass. In this first pass, she considers outcome data on fish biomass collected only at treated sites since the intervention began. Based on this data, no quasi-experimental methods are applicable.

STEP 2A: using Decision Matrix 2 to select quasi-experimental methods - second pass

		Decision Matrix 2		
		1. What outcome data do you have and/or can you get?		
		Pre- & post-intervention time series	Pre- & post-intervention data (single observation or time-series)	Post-intervention data only (single observation or time-series)
2. You have this data for…	Treated & non-treated sites	All methods possible	Matching Difference-in-differences Regression discontinuity	Matching (single difference analysis) Regression discontinuity
	Treated sites only	Interrupted time series	No methods possible (*revisit Step 1: Decision Matrix 1)*	No methods possible *(revisit Step 1: Decision Matrix 1)*

Step 2A: evaluator revisits Decision Matrix 2 – in this second pass, she considers time-series fish catch data collected at treated and non-treated sites, which she will use as a proxy for fish biomass (the outcome of interest). She can use all methods with this data.

FIGURE 3 Illustrating use of Decision Matrix 1 and 2 to select counterfactual methods to evaluate a community-based marine management project. Blue shaded cells indicate options selected by evaluator after answering questions. Textboxes on the right explain how the evaluator navigated the decision matrices, iterating twice over Decision Matrix 2 for different outcome data.

intervention single observation or time-series" in Decision Matrix 2 (see Figure 3). However, this data is available for "treated sites only" (question 2 in Decision Matrix 2), so *no options are possible with this data*. This reflects a common challenge in conservation: substantial effort is often invested in data collection without consideration of evaluation needs, resulting in data that cannot support meaningful analysis.

The evaluator however realizes there is national time-series fish catch data at the municipal level spanning 10 years before and 5 years after the intervention, which could be a potential proxy for fish biomass. After discussing this option with the project team and gaining consensus, the evaluator revisits Decision Matrix 2, confirming that this data could be used to estimate impact using all possible quasi-experimental methods.

Proceeding to Step 3, the evaluator engages a researcher specializing in counterfactual methods to assess the minimum sample size required to statistically detect an effect of the intervention. Since the intervention is applied at the municipal (water) level, municipalities are the primary unit of analysis. A power analysis suggests that a minimum sample size of $n = 284$ (142 intervention and 142 control sites) is necessary to detect an existing effect. Government records indicate there are approximately 800 coastal municipalities in the region (including the 160 intervention sites); this should be sufficient for selecting suitable control sites that can represent the counterfactual.

Finally, in (Step 4), the evaluator formalizes the collaboration with the researcher who can provide the technical expertise needed for the evaluation and can provide capacity-building for future evaluation. The evaluator also confirms she has the budget to cover labor and software costs, so she decides to proceed with implementing the counterfactual impact evaluation. As a final note, given that the data allows any quasi-experimental approach to be used, the evaluation team consults Table 1 and decides to implement DID as it is more intuitive for partners and other stakeholders. Before proceeding, they check whether fish catch data at intervention and non-intervention sites follow parallel trends prior to the intervention; this is needed for use of DID (see Table 1). They confirm that the pre-intervention trends are parallel and proceed with DID.

5 | LOOKING AHEAD

As awareness about the importance of evidence-based decision-making in conservation is increasing (e.g., Pullin & Knight, 2003), so is the demand for robust causal methods to generate this evidence (Ferraro, 2009; Miteva et al., 2012). Our aim here is to support this transformation by presenting practical guidance for conservation practitioners, researchers, and funders seeking to integrate counterfactual impact evaluation methods into their work. The aim is to help budding evaluators assess the feasibility of different impact evaluation approaches for specific projects. It also aims to outline the requirements for conducting impact evaluations, encouraging the integration of these considerations into future project planning.

While we provide guidance on method selection, detailed implementation procedures are outside the scope of this paper. Implementation will require differing levels of technical skill, depending on the method used, as well as a deep understanding of the intervention that is being evaluated. Many of the decisions and assumptions made when using this framework will need to be revisited, including considerations about: unobservable confounders and their relative effect(s) on treatment assignment; potential spillover effects and how to mitigate these; timescales for identifying intervention effects; analytical methods for clustered and/or time-series data.

Looking ahead, the expectation is that counterfactual methods will be integrated into conservation practice, guiding project design and implementation—particularly for interventions that are: unproven, high stakes, expensive, large in geographic scope, and/or which require evidence of additionality (Neugarten et al., 2025). This may require restructuring the traditional conservation project cycle, where evaluation becomes a seamless part of project design and implementation rather than a separate phase. This integration is expected to become easier with advancements in geospatial data collection, bolstered by AI and machine learning. Achieving this integration also depends on positive incentives for impact evaluations (Asquith, 2020) and embracing a "safe to fail culture" in conservation (Catalano et al., 2018, 2019). Funders play a crucial role by ensuring long-term funding for implementers who use evaluation evidence in decision-making. Together, these efforts will be critical in promoting evidence-based practices in conservation that benefit both biodiversity and people.

AUTHOR CONTRIBUTIONS

All authors contributed to the conceptual design, drafting, and writing of the paper.

ACKNOWLEDGMENTS

We thank Sebastien Costedoat from Conservation International for providing valuable input for this paper. RM acknowledges the Conservation X Labs team for their time and insight, including Henrik Cox, Daphne Yin, and the Co-Founders Alex Dehgan and Paul Bunje for their commitment to evaluation and learning. JE acknowledges funding from the Academy of Finland (grant no. 333518) and the Kone foundation. Open access publishing facilitated by Helsingin yliopisto, as part of the Wiley - FinELib agreement.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

There is no data associated with this paper.

ORCID

Tanya O'Garra https://orcid.org/0000-0002-8214-1872
Rachel Martin https://orcid.org/0000-0001-5914-2132
Claudia Polo-Urrea https://orcid.org/0009-0001-1834-3250
Johanna Eklund https://orcid.org/0000-0003-1263-8151

ENDNOTES

[1] Although identifying and controlling for confounders is not essential for the canonical versions of some of the methods presented here (e.g., RCTs, RDD, and SCM), in practice they are highly recommended (Abadie & Vives-i-Bastida, 2022; Cattaneo et al., 2023) as they help to reduce the standard error (hence increase power, see Section 3, Part 3) and reduce bias (Schaffer et al., 2021). For some methods (specifically, RDD, see Table 1), however, the current recommendation—which we subscribe to with respect to all methods used—is to present estimates of impact with and without adjusting for covariates side by side (Cattaneo et al., 2023).

[2] The counterfactual methods presented here can be used to identify the effect of specific mechanisms in the causal chain; in this case, the mechanism under study can be considered the "intervention". However, these methods are not in themselves sufficient to fully analyze entire causal chains (Ferraro & Hanauer, 2014).

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*, 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control methods. *American Journal of Political Science*, *59*, 495–510.

Abadie, A., & Vives-i-Bastida, J. (2022). Synthetic controls in action. arxiv.org/abs/2203.06279.

Adams, V. M., Barnes, M., & Pressey, R. L. (2019). Shortfalls in conservation evidence: Moving from ecological effects of interventions to policy evaluation. *One Earth*, *1*(1), 62–75.

Aldashev, G., Kirchsteiger, G., & Sebald, A. (2017). Assignment procedure biases in randomised policy experiments. *The Economic Journal*, *127*(602), 873–895.

Alix-Garcia, J. M., Sims, K. R. E., Orozco-Olvera, V. H., & Romo Monroy, S. (2018). Payments for environmental services supported social capital while increasing land management. *Proceedings of the National Academy of Sciences*, *115*(27), 7016–7021. https://doi.org/10.1073/pnas.1720873115

Andam, K. S., Ferraro, P. J., Pfaff, A., & Robalino, J. A. (2008). Measuring the effectiveness of protected area networks in reducing deforestation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(42), 16089–16094.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, *111*(12), 4088–4118.

Asquith, N. (2020). Large-scale randomized control trials of incentive-based conservation: What have we learned? *World Development*, *127*, 104785.

Baragwanath, K., & Bayi, E. (2020). Collective property rights reduce deforestation in the Brazilian Amazon. *Proceedings of the National Academy of Sciences*, *117*(34), 20495–20502.

Baylis, K., Honey-Rosés, J., Börner, J., Corbera, E., Ezzine-de-Blas, D., Ferraro, P. J., Lapeyre, R., Persson, U. M., Pfaff, A., & Wunder, S. (2016). Mainstreaming impact evaluation in nature conservation. *Conservation Letters*, *9*(1), 58–64.

Bell, M. L., Kenward, M. G., Fairclough, D. L., & Horton, N. J. (2013). Differential dropout and bias in randomized controlled trials: When it matters and when it may not. *BMJ (Clinical Research Ed.)*, *2013*, e8668.

Blake, K., Kubo, T., & Veríssimo, D. (2023). Measuring the effectiveness of value-framing and message valence on audience engagement across countries. *Global Environmental Psychology*, *1*, e11181.

Bull, J. W., Strange, N., Smith, R. J., & Gordon, A. (2021). Reconciling multiple counterfactuals when evaluating biodiversity conservation impact in social–ecological systems. *Conservation Biology*, *35*(2), 510–521.

Bulte, E., Beekman, G., Di Falco, S., Hella, J., & Lei, P. (2014). Behavioral responses and the impact of new agricultural technologies: Evidence from a double-blind field experiment in Tanzania. *American Journal of Agricultural Economics*, *96*(3), 813–830.

Catalano, A. S., Redford, K., Margoluis, R., & Knight, A. T. (2018). Black swans, cognition, and the power of learning from failure. *Conservation Biology*, *32*(3), 584–596.

Catalano, A. S., Lyons-White, J., Mills, M. M., & Knight, A. T. (2019). Learning from published project failures in conservation. *Biological Conservation*, *238*, 108223.

Cattaneo, M. D., Keele, L., & Titiunik, R. (2023). Covariate adjustment in regression discontinuity designs. Handbook of matching and weighting adjustments for causal inference, pp. 153–168.

Conservation X Labs. (n.d.). Conservation X Labs. Retrieved January 3, 2025, from https://conservationxlabs.com/

Craigie, I. D., Barnes, M. D., Geldmann, J., & Woodley, S. (2015). International funding agencies: Potential leaders of impact evaluation in protected areas? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*, 20140283.

Curzon, H. F., & Kontoleon, A. (2016). From ignorance to evidence? The use of programme evaluation in conservation: Evidence from a Delphi survey of conservation experts. *Journal of Environmental Management*, *180*, 466–475.

Dee, L. E., Ferraro, P. J., Severen, C. N., Kimmel, K. A., Borer, E. T., Byrnes, J. E. K., Clark, A. T., Hautier, Y., Hector, A., Raynaud, X., Reich, P. B., Wright, A. J., Arnillas, C. A., Davies, K. F., MacDougall, A., Mori, A. S., Smith, M. D., Adler, P. B., Bakker, J. D., ... Loreau, M. (2023). Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nature Communications*, *14*(2607), 1–12.

Devenish, K., Desbureaux, S., Willcock, S., & Jones, J. P. G. (2022). On track to achieve no net loss of forest at Madagascar's biggest mine. *Nature Sustainability*, *5*, 498–508.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Duflo, E., Glennerster, R., & Kremer, M. (2008). Chapter 61 using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Dunn, M. E., Mills, M., & Veríssimo, D. (2020). Evaluating the impact of the documentary series blue planet II on viewers' plastic consumption behaviors. *Conservation Science and Practice*, *2*(10), e280.

Feng, Y., Wang, Y., Su, H., Pan, J., Sun, Y., Zhu, J., Fang, J., & Tang, Z. (2021). Assessing the effectiveness of global protected areas based on the difference in differences model. *Ecological Indicators*, *130*, 108078.

Ferraro, P. J. (2009). Counterfactual thinking and impact evaluation in environmental policy. *New Directions for Evaluation*, *122*, 75–84.

Ferraro, P. J. (2012). *Experimental project designs in the global environment facility: Designing projects to create evidence and catalyze investments to secure global environmental benefits. A STAP advisory document*. Global Environment Facility.

Ferraro, P. J., & Hanauer, M. M. (2014). Advances in measuring the environmental and social impacts of environmental programs. *Annual Review of Environment and Resources*, *39*, 495–517.

Ferraro, P. J., & Hanauer, M. M. (2015). Through what mechanisms do protected areas affect environmental and social outcomes? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1681), 20140267.

Ferraro, P. J., Hanauer, M. M., & Sims, K. R. E. (2011). Conditions associated with protected area success in conservation and poverty reduction. *PNAS*, *108*(34), 13913–13918.

Ferraro, P. J., & Pattanayak, S. K. (2006). Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *PLoS Biology*, *4*, e105.

Ferraro, P. J., & Pressey, R. L. (2015). Measuring the difference made by conservation initiatives: Protected ateas and their environmental and social impacts. *Philosophical Transactions of the Royal Society B*, *370*, 201400270.

Ferraro, P. J., Sanchirico, J. N., & Smith, M. D. (2019). Causal inference in coupled human and natural systems. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(12), 5311–5318.

Fives, A., Russell, D. W., Canavan, J., Lyons, R., Eaton, P., Devaney, C., Kearns, N., & O'Brien, A. (2013). The ethics of randomized controlled trials in social settings: Can social trials be scientifically promising and must there be equipoise? *International Journal of Research & Method in Education*, *38*(1), 56–71.

Fox, A. K., Molina, A. C., & Swearingen, T. C. (2022). An interrupted time series approach to assess marine protected area impacts on recreational fishing licence sales. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *32*(12), 1970–1982.

Francini-Filho, R. B., & Moura, R. L. (2008). Evidence for spillover of reef fishes from a no-take marine reserve: An evaluation using the before-after control-impact (BACI) approach. *Fisheries Research*, *93*(3), 346–356.

Fredriksson, A., & de Magalhães Oliveira, G. (2019). Impact evaluation using difference-in-differences. *RAUSP Management Journal*, *54*(4), 519–532.

GEF IEO. (2016). Impact Evaluation of GEF Support to Protected Areas and Protected Area Systems, Evaluation Report No. 104. Global Environment Facility Independent Evaluation Office, Washington, DC: GEF IEO.

Gertler, P. J., Martinez, S., Premand, P. L., Rawlings, L. V., & Vermeersh, C. M. J. (2016). *Impact evaluation in practice* (second ed.). The World Bank.

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton University Press.

Greenfield, P. (2023). More than 90% of rainforest carbon offsets by biggest certifier are worthless, analysis shows. The Guardian, 18 Jan.

Greenstone, M., & Gayer, T. (2009). Quasi-experimental and experimental approaches to environmental economics. *Journal of Environmental Economics and Management*, *57*(1), 21–44.

Gregory-Smith, D., Wells, V. K., Manika, D., & McElroy, D. J. (2017). An environmental social marketing intervention in cultural heritage tourism: A realist evaluation. *Journal of Sustainable Tourism*, *25*(7), 1042–1059.

Grillos, T., Bottazzi, P., Crespo, D., Asquith, N., & Jones, J. P. G. (2019). In-kind conservation payments crowd in environmental values and increase support for government intervention: A randomized trial in Bolivia. *Ecological Economics*, *166*, 106404.

Gupta, S., Kopper, S., Cavanagh, J., Doyle, M. A., Duru, M., Feeney, L., Gibson, M., & Naimpally, R. (2022). Power calculations. JPAL Resources. https://www.povertyactionlab.org/resource/power-calculations

Jayachandran, S., de Laat, J., Lambin, E. F., Stanton, C. Y., Audy, R., & Thomas, N. E. (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science*, *357*(6348), 267–273.

Jones, I. J., MacDonald, A. J., Hopkins, S. R., & Sokolow, S. H. (2020). Improving rural health care reduces illegal logging and conserves carbon in a tropical forest. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(45), 28515–28524.

Joppa, L. N., & Pfaff, A. (2009). High and far: Biases in the location of protected areas. *PLoS One*, *4*(12), e8273.

Kellogg, M., Mogstad, M., Pouliot, G. A., & Torgovitsky, A. (2019). Combining matching and synthetic control to trade off biases from extrapolation and interpolation. *Journal of the American Statistical Association*, *116*(536), 1804–1816.

Kimmel, K., Dee, L. E., Avolio, M. L., & Ferraro, P. J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, *36*(12), 1141–1152.

Kleiman, D. G., Reading, R. P., Miller, B. J., Clark, T. W., Scott, J. M., Robinson, J., Wallace, R. L., Cabin, R. J., & Felleman, F. (2000). Improving the evaluation of conservation programs. *Conservation Biology*, *14*(2), 356–365.

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., & Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, *25*(12), 1514–1528.

Langhammer, P. F., Bull, J. W., Bicknell, J. E., Oakley, J. L., Brown, M. H., Bruford, M. W., Butchart, S. H. M., Carr, J. A., Church, D., Cooney, R., Cutajar, S., Foden, W., Foster, M. N., Gascon, C., Geldmann, J., Genovesi, P., Hoffmann, M., Howard-McCombe, J., Lewis, T., ... Brooks, T. M. (2024). The positive impact of conservation action. *Science*, *384*(6694), 453–458.

Linden, A. (2018). A matching framework to improve causal inference in interrupted time-series analysis. *Journal of Evaluation in Clinical Practice*, *24*(2), 408–415.

Mahajan, S. L., Tanner, L., Ahmadia, G., Becker, H., DeMello, N., Fidler, R., Harborne, A. R., Jagadish, A., Mills, M., Cairney, P., Cheng, S., Fariss, B., Masuda, Y. J., Pabari, M., Tengö, M., Wyborn, C., & Glew, L. (2023). Accelerating evidence-informed decision-making in conservation implementing agencies through effective monitoring, evaluation, and learning. *Biological Conservation*, *286*, 110304.

Martin, A., Gross-Camp, N. D., Kebede, B., & McGuire, S. (2014). Measuring effectiveness, efficiency and equity in an

experimental payments for ecosystem services trial. *Global Environmental Change*, *28*(1), 216–226.

McKenzie, D. (2023). What to do about parallel trends when you only have baseline data? In *Development Impact*. World Bank. https://blogs.worldbank.org/en/impactevaluations/what-do-about-parallel-trends-when-you-only-have-baseline-data

McKinnon, M. C., Mascia, M. B., Yang, W., Turner, W. R., & Bonham, C. (2015). Impact evaluation to communicate and improve conservation non-governmental organization performance: The case of Conservation International. *Philosophical Transactions of the Royal Society B*, *370*(1681), 20140282.

Miteva, D. A., Pattanayak, S. K., & Ferraro, P. J. (2012). Evaluation of biodiversity policy instruments: What works and what doesn't? *Oxford Review of Economic Policy*, *28*, 69–92.

Molina-Millán, T., & Macours, K. (2025). Attrition in randomized controlled trials: Using tracking information to correct bias. *Economic Development and Cultural Change*, *73*(2), 811–834.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference methods and principles for social research*. Cambridge University Press.

Neugarten, R., Rodenwald, A., Eklund, J., & O'Garra, T. (2025). An introduction to impact evaluation for conservation. *Conservation Science and Practice*.

O'Garra, T., Mangubhai, S., Jagadish, A., Tabunakawai-Vakalalabure, M., Tawake, A., Govan, H., & Mills, M. (2023). National-level evaluation of a community-based marine management initiative. *Nature Sustainability*, *6*, 908–918.

Ota, M., Ota, T., Shimizu, K., Onda, N., Ma, V., Sokh, H., & Mizoue, N. (2023). Forest conservation effectiveness of community forests may decline in the future: Evidence from Cambodia. *PNAS Nexus*, *2*(10), 320.

Peters, J., Langbein, J., & Gareth, R. (2018). Generalization in the tropics—development policy, randomized controlled trials, and external validity. *The World Bank Research Observer*, *33*(1), 34–64.

Plümper, T., Troeger, V. E., & Neumayer, E. (2019). Case selection and causal inferences in qualitative comparative research. *PLoS One*, *14*(7), e0219727.

Pullin, A. S., & Knight, T. M. (2003). Support for decision making in conservation practice: An evidence-based approach. *Journal for Nature Conservation*, *11*(2), 83–90.

Puri, J., Rastogi, A., Prowse, M., & Asfaw, S. (2020). Commentary. Good will hunting: Challenges of theory-based impact evaluations for climate investments in a multilateral setting. *World Development*, *127*, 104784.

Pynegar, E., Booth, H., Doulton, H., Ferraro, P. J., Mohamed, M., Rakotonarivo, O. S., & Jones, J. P. (2025). RCTs in the wild: Designing and implementing conservation programs as randomized control trials. *Conservation Science and Practice*, e70029.

Pynegar, E. L., Gibbons, J. M., Asquith, N. M., & Jones, J. P. G. (2021). What role should randomized control trials play in providing the evidence base for conservation? *Oryx*, *55*(2), 235–244.

Pynegar, E. L., Jones, J. P. G., Gibbons, J. M., & Asquith, N. M. (2018). The effectiveness of payments for ecosystem services at delivering improvements in water quality: Lessons for experiments at the landscape scale. *PeerJ*, *6*, e5753.

Rana, P., & Miller, D. C. (2021). Predicting the long-term social and ecological impacts of tree-planting programs: Evidence from northern India. *World Development*, *140*, 105367.

Reimer, M. N., & Haynie, A. C. (2018). Mechanisms matter for evaluating the economic impacts of marine reserves. *Journal of Environmental Economics and Management*, *88*, 427–446.

Robalino, J., Pfaff, A., & Villalobos, L. (2017). Heterogeneous local spillovers from protected areas in Costa Rica. *Journal of the Association of Environmental Economists*, *4*(3), 795–820.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1980). Comment on: "Randomization analysis of experimental data in the fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, *75*, 591–593.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decision. *Journal of the American Statistical Association*, *100*(469), 322–331.

Sabin, S., Dieudonne, B., Mitchell, J., White, J., Chin, C., & Morikawa, R. (2019). Community-based watershed change: A case study in eastern Congo. *Forests*, *10*(6), 475.

Salazar, G., Mills, M., & Veríssimo, D. (2018). Qualitative impact evaluation of a social marketing campaign for conservation. *Conservation Biology*, *33*(3), 634–644.

Schaffer, A. L., Dobbins, T. A., & Pearson, S. A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*, *21*(1), 58.

Schleicher, J., Eklund, J., Barnes, M. D., Geldmann, J., Oldekop, J. A., & Jones, J. P. G. (2020). Statistical matching for conservation science. *Conservation Biology*, *34*, 538–549.

Sharma, R., Jones, S., Robinson, D., & Gordon, A. (2023). Evaluating the impact of private land conservation with synthetic control design. *Conservation Biology*, *37*(6), e14150.

Sills, E. O., Herrera, D., Kirkpatrick, A. J., Brandão Jr, A., Dickson, R., Hall, S., Pattanayak, S., Shoch, D., Vedoveto, M., Young, L., & Pfaff, A. (2015). Estimating the impacts of local policy innovation: The synthetic control method applied to tropical deforestation. *PLoS One*, *10*(7), 0132590.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Stuart, E. A., & Lalongo, N. S. (2010). Matching methods for selection of participants for follow-up. *Multivariate Behavioral Research*, *45*(4), 746–765.

Taskforce on Nature-related Financial Disclosures. (2023). Recommendations of the Taskforce on nature-related financial disclosures. https://tnfd.global/wp-content/uploads/2023/08/Recommendations_of_the_Taskforce_on_Nature-related_Financial_Disclosures_September_2023.pdf?v=1695118661

Treurniet, M. (2023). The impact of being surveyed on the adoption of agricultural technology. *Economic Development and Cultural Change*, *71*(3), 941–962.

Turner, S. L., Karahalios, A., Forbes, A. B., Taljaard, M., Grimshaw, J. M., & McKenzie, J. E. (2021). Comparison of six statistical methods for interrupted time series studies: Empirical evaluation of 190 published series. *BMC Medical Research Methodology*, *21*, 134.

Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation. *American Economic Review*, *105*(5), 467–470.

Wauchope, H. S., Amano, T., Geldmann, J., Johnston, A., Simmons, B. I., Sutherland, W. J., & Jones, J. P. G. (2021). Evaluating impact using time-series data. *Trends in Ecology & Evolution*, *36*(3), 196–205.

Wauchope, H. S., Jones, J. P. G., Geldmann, J., Simmons, B. I., Amano, T., Blanco, D. E., Fuller, R. A., Johnston, A., Langendoen, T., Mundkur, T., Nagy, S., & Sutherland, W. J. (2022). Protected areas have a mixed impact on waterbirds, but management helps. *Nature*, *605*, 103–107.

West, T. A. P., Börner, J., Sills, E. O., & Kontoleon, A. (2020). Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon. *PNAS*, *117*(39), 24188–24194.

West, A. P., Wunder, S., Sills, E. O., Börner, J., Rifai, S. W., Neidermeier, A. N., Frey, G. P., & Kontoleon, A. (2023). Action needed to make carbon offsets from forest conservation work for climate change mitigation. *Science*, *381*, 873–877.

Wiik, E., Jones, J. P. G., Pynegar, E., Bottazzi, P., Asquith, N., Gibbons, J., & Kontoleon, A. (2020). Mechanisms and Impacts of an Incentive-Based Conservation Program with Evidence from a Randomized Control Trial. *Conservative Biology*, *34*(5), 1076–1088.

World Economic Forum. (2020). The future of nature and business. New Nature Economy Report II. The Future of Nature and Business. https://www3.weforum.org/docs/WEF_The_Future_Of_Nature_And_Business_2020.pdf

Wuepper, D., & Finger, R. (2023). Regression discontinuity designs in agricultural and environmental economics. *European Review of Agricultural Economics*, *50*(1), 1–28.

Zavaleta Cheek, J., Eklund, J., Merten, N., Brooks, J., & Miller, D. C. (2023). A guide to qualitative attribution methods for evaluation in conservation. *Conservation Biology*, *37*(4), e14071. https://doi.org/10.1111/cobi.14071

Zwane, A. P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., Kremer, M., Karlan, D. S., Hornbeck, R., & Giné, X. (2011). Being surveyed can change later behavior and related parameter estimates. *Proceedings of the National Academy Sciences*, *108*(5), 1821–1826. https://doi.org/10.1073/pnas.1000776108

SUPPORTING INFORMATION

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

How to cite this article: O'Garra, T., Martin, R., Pynegar, E., Polo-Urrea, C., & Eklund, J. (2025). Selecting among counterfactual methods to evaluate conservation interventions. *Conservation Science and Practice*, e70066. https://doi.org/10.1111/csp2.70066