RCTs in the Wild: Designing and implementing conservation programs as Randomized Control Trials

Edwin Pynegar[1,2], Hollie Booth[2,3], Hugh Doulton[4], Paul J. Ferraro[5], Misbahou Mohamad[4], O. Sarobidy Rakotonarivo[6], Julia P G Jones[1,7*]

[1] School of Environmental and Natural Sciences, Bangor University

[2] The Biodiversity Consultancy

[3] Department of Biology, University of Oxford

[4]Dahari, Mutsumudu, The Union of the Comoros

[5] Carey Business School and the Department of Environmental Health and Engineering, Johns Hopkins University

[6] Ecole Supérieure des Sciences Agronomiques, University of Antananarivo

[7] Department of Biology, Utrecht University

*Corresponding author: julia.jones@bangor.ac.uk

Abstract

Experimental evaluation of environmental programs, including those aiming to conserve biodiversity, are rarer than those of social programs. However, there is growing interest in conducting such evaluations in conservation. Randomized Control Trials (RCTs), in which units are randomly assigned to receive one of two or more treatments, can avoid biases associated with observational designs and provide reliable estimation of program effectiveness. We present a typology of conservation RCTs, which differentiates between interventions that have a direct impact on biodiversity outcomes and those where the impact is mediated through changes in human behavior. With a focus on RCTs in behaviorally mediated conservation programs, which have received limited attention, we examine: 1) technical challenges (selection of randomization unit, avoiding interference between units, ensuring excludability and external validity) and solutions, 2) ethical and practical challenges and 3) incentives that conservation organizations have to run RCTs. We end by summarizing the steps needed for good design and transparent reporting of RCTs. RCTs are not always appropriate, but we believe conservation science and practice would benefit from their wider application. By demystifying RCTs of conservation programs we hope that this article will serve as a practical, grounded guide to foster broader adoption.

Introduction

Biodiversity conservation is going through a causal revolution, with a rapid expansion in the use of better methods for understanding causes and effects (Jones & Shreedhar, 2024). Such understanding is crucial to allow accurate predictions of the impacts of interventions in real-world settings (Ferraro, et al., 2023) and therefore to advance the development of effective conservation programs and policies. Most conservation programs are behaviorally mediated, in the sense that outcomes for biodiversity are achieved through changing human behavior. The design and implementation of robust evaluations of behaviorally mediated programs in coupled human-natural systems presents particular challenges (Ferraro et al., 2019) However, while quasi-experimental approaches to evaluating such programs have seen relatively widespread uptake, experimental approaches are rarer (Ferraro & Messer this issue; Alpízar & Ferraro, 2020; Behaghel et al., 2019; Ma et al., 2017; Pynegar et al., 2021).

Randomized control trials (RCTs), where units are randomly assigned to two or more treatments, avoid many of the challenges of causal inference from quasi-experimental approaches (Ferraro & Hanauer, 2014). In the context of behaviorally mediated conservation programs, RCTs make it possible to isolate the impact of an intervention separate from the many unobservable drivers of human decisions that also affect conservation outcomes (Ferraro and Messer this issue). This is because units allocated to the program should be similar to those not allocated to the program in terms of these confounding variables, and therefore would have the same expected outcomes in the absence of the program (Ferraro, 2012). In other words, those not allocated to the program, the control group, represent a valid counterfactual (Ferraro, 2009). Comparing post-intervention outcomes in treatment and control groups provides a valid estimate of the impact of the program. As well as comparing the impact of a program relative to no program, RCTs can be used to explore the effects of different versions of program (Wardropper et al., 2022).

While there have been important criticisms of RCTs for program evaluation (Barrett, 2021; Deaton & Cartwright, 2018; Ravallion, 2020), the use of RCTs has revolutionized several areas of policy evaluation where the effectiveness of the intervention depends on how people respond to the intervention, such as in the contexts of rural development (Banerjee et al., 2020; Webber & Prouse, 2018), public health (Jones & Podolsky, 2015), and education (Styles & Torgerson, 2018). Yet, the substantial literature on RCT design for non-environmental program evaluations (Connolly et al., 2017; Glennerster & Takavarasha, 2013; Matthews, 2006) may not be easily accessible to conservation organizations that are considering how best to use RCTs in their programs.

Our aim is to contribute to accelerating the use of RCTs in conservation programs, thereby increasing understanding of what works in conservation. We first present a typology of conservation RCTs, which distinguishes between interventions which have a direct impact on biodiversity and those where impacts on biodiversity are mediated through changes in human behavior. The use of RCTs in behaviorally mediated programs has received much less attention than the use of RCTs in conservation interventions with more direct impact on biodiversity (see, for example, Smith et al., 2023). After presenting the typology, we explore the challenges to greater application of RCTs in behaviorally mediated programs, and describe possible solutions to these challenges, by drawing on our own experiences as implementers of conservation RCTs (Figure 1), as well as the wider literature. We end by summarizes the steps needed for good design and reporting of RCTs.

A typology of conservation RCTs

Research questions suitable for answering using an RCT are often framed using a PICO framework (Population, Intervention, Comparator, Outcome; e.g. Huang et al., 2006) (Figure 1).

The **population** in an RCT is the groups of units about which the research seeks to be able to generalize to. In most RCTs, the units are people, such as individuals with a specific condition in a medical trial, or with particular socio-economic characteristics in development, public health or education trials. However, in conservation RCTs, the units often need to be given careful thought because outcomes such as deforestation or offtake of wild species need to be associated with a defined area of space, while other outcomes such as wellbeing will be associated with individuals or households (Figure 1).

Interventions in conservation can directly impact biodiversity by, for example, controlling invasive species, excluding grazing from an area to support natural regeneration, or providing artificial nest holes for threatened species. The Conservation Evidence database provides an invaluable resource summarizing evidence concerning this sort of intervention (Sutherland & Wordley, 2018). However, many conservation programs seek to impact biodiversity by changing human behavior (Jones & Shreedhar, 2024; Nielsen et al., 2021). For example, protected areas or species conservation rules, incentive-based measures, training, and nudges all involve behavior as a mediator between the intervention and biodiversity outcomes (Byerly et al., 2018).

The **comparator** is the group or groups to which the treatment group will be compared (sometimes called 'control' or 'comparison' group, condition or arm). In a traditional two-arm RCT there is a single treatment group and a control group. However larger RCTs may include multiple treatment arms as well as a control group.

The ultimate **outcomes** of interest in many conservation RCTs are ecological (such as changes in ecosystem extent or condition, species abundance or diversity; Table 1 left hand side). However, social outcomes may also be important for several reasons (Table 1 right hand side). Firstly, social metrics – such as behaviors, attitudes and norms - can act as proxies for impacts on biodiversity when linked through a clear and testable theory of change (Cheng et al., 2020; Jones & Shreedhar, 2024), and these proxies are often easier to measure and more sensitive to change than biodiversity outcomes. Secondly, conservation practitioners increasingly seek to understand the social impacts of interventions on human wellbeing, or perception of equity. They seek such understanding for both moral reasons (to ensure conservation at a minimum does no harm) and instrumental reasons (conservation is typically more successful when perceived positively locally (Oyanedel et al., 2020).

The vast majority of conservation RCTs conducted to date focus on interventions with direct impact on ecological outcomes (Ockendon et al., 2021), such excluding predators from important sites for wading birds or eradicating rats from islands. While more such experimental evaluations are certainly needed (Smith et al., 2023), they have created a valuable evidence base for wildlife and land managers (Christie et al., 2020; Conservation Evidence; Table 1 top left box). However, many conservation programs involve some degree of behavioral mediation (for example incentivizing farmers to change land management practices, or patrolling aiming to increase compliance with conservation rules). RCTs of such behaviorally mediated conservation programs (Table 1 bottom row) are the focus of the remainder of this article.

Challenges and solutions in RCTs of behaviorally mediated conservation programs

1) Technical challenges

Coupled human-natural systems are complex, dynamic and have feedbacks between social and environmental components which poses particular challenges for evaluating conservation programs (Ferraro et al., 2019). It is important to note that while we emphasize how these challenges affect RCTs, most of these challenges are shared in some form by other impact evaluation designs. We first address the challenge of selecting the appropriate unit for an RCT and how that interacts with the power of an RCT to detect a treatment effect. We then explore two key assumptions which must be met for an RCT to provide an accurate estimate of the treatment effect (i.e. to have internal validity): no interference between units (i.e. no spillovers) and excludability (the randomization process can only influence outcomes through its effect on treatment status). Finally, we discuss challenges to ensure an RCT has external validity (the results can be generalized). For each challenge we also highlight possible solutions (Table 2).

Selection of randomization unit

Identifying the randomization unit for a conservation RCT involves considering the scale at which the program is targeted (e.g., individuals, villages, watersheds), the scale at which outcomes can be measured (e.g., individuals, landscapes, watersheds, species ranges; Figure 2), as well as potential spillovers (Figure 3) and the implications of the scale for statistical power. In most RCTs, units are randomized as individuals (e.g., patients in a medical trial are randomly allocated to receive a new treatment or not) or by clusters of individuals such as households, administrative areas, or schools. Cluster RCTs may be done because the intervention is most conveniently delivered via clusters (e.g. everyone in a school receives an education intervention), because unequal treatment of individuals within a cluster may cause dissatisfaction, because the outcome is most sensibly measured at the level of the cluster, or because units within the cluster may interfere with each other (i.e., there may be spillovers between units (e.g. Donner & Klar, 2004; Gertler et al., 2016; see next subsection for details).

Many outcomes in conservation RCTs have a spatial element, meaning boundaries of areas to be used as randomization units are required. For example, the community may be the appropriate randomization unit in an RCT that evaluates payments to reduce deforestation when forests are not individually owned (Jayachandran et al., 2016; Wiik et al., 2019; Wilebore et al., 2019). However, measuring deforestation at the community scale requires units for which there are recognized spatial boundaries. Jayachandran et al., (2016)were able to access village boundaries provided by the Ugandan Bureau of Statistics. Wiik et al., (2019) delineated boundaries between communities based upon land tenure documentation held by participants, combined with key informant interviews and reference to topographical features. Wilebore et al., (2019) produced polygons with boundaries equidistant between village centers, which were subsequently improved through ground truthing (Wilebore & Coomes, 2016).

The number of units which can be randomized has a strong influence on the power of an RCT to detect an impact from the program when one exists. Running an underpowered RCT can do more harm than good because its wastes resources and will tarnish the reputation of a potentially effective program. In technical terms, the power of an RCT is the probability that, for a given effect size and a given statistical significance level, the hypothesis of zero effect can be rejected (Duflo et al., 2007). The minimal detectable effect of an RCT therefore depends on the number of units and the variability in the outcome e.g. Gertler et al., 2016; Glennerster & Takavarasha, 2013). Selecting the smallest feasible randomization unit for the intervention will tend to increase the sample size and thus maximize the minimal detectable effect given resources available. For this reason, it would be difficult to run an RCT to explore the impact of an intervention on the relative abundance of a widely ranging species, because the resource requirements of monitoring and logistics over sufficient units would likely be prohibitive (Figure 2). In cluster RCTs, power depends upon the number of clusters, the number of units within each cluster, and the intra-cluster correlation or degree of similarity of outcomes of the units within each cluster over and above those in other clusters. Similarity within clusters will lead to a power closer to that of the number of clusters; lack of similarity will result in a power closer to that of the number of units (e.g. Gertler et al., 2016).

Avoiding interference between units

In RCTs, an important assumption is that the outcome of a unit depends only on its own intervention status. When this assumption is violated, units are said to "interfere" with each other and this interference complicates the interpretation of results from an RCT. Interference in coupled human-natural systems can occur in three main ways (Ferraro et al., 2019, Figure 3)

Firstly, an intervention that aims to reduce environmentally-damaging behaviors may displace people from treated areas to control areas, where they continue with environmentally-damaging activities (type i spillover *sensu* Ferraro et al., 2019). Secondly, in response to reductions in environmentally damaging activities in the treated area, people in control areas may increasing the intensity of their activities (type ii spillover). For example in the Dahari RCT (Figure 3), landowners incentivized to cease deforestation may move their activities into areas used by control-group landowners (Figure 3i) or control-group landowners may increase deforestation on land to meet demand (Figure 3ii). Both spillovers tend to inflate estimates of an intervention's impact meaning that an intervention may appear to be effective when there has been no impact on outcomes at the landscape scale because pressures have simply been displaced from treatment to control-associated areas. The term leakage is commonly applied to describe these sorts of spillovers in the context of conservation programs (Meyfroidt et al., 2020).

Thirdly, the treatment may induce change in outcomes outside of the treated units through material flows, such as dispersal of individuals of a species or flows of water (type iii spillovers). This type of spillover will tend to reduce apparent effect sizes. For example, if the payments in the Dahari RCT resulted in benefits to habitat (less deforestation) and therefore to threatened bats, these outcomes might spillover into control units, reducing the estimated impact of the intervention (Figure 3iii).

Similar phenomenon can apply even if the intervention is not area-based. For example in the KUL RCT, fishers in the control group may increase the number of sharks and wedgefish they sell if the price rises as a result of the reduction in supply in the treated group. This would be an example of a type ii spillover and would inflate the estimated impact of the intervention.

There are RCT designs which can be used to quantify interference (Baird et al., 2017). For example, randomizing clusters to treatment or control and then randomizing units within the treatment clusters and comparing outcomes for control units in treatment and control clusters can give an estimate of the magnitude of spillovers within clusters (assuming there is not spillover between clusters). However, such designs are often difficult when an RCT has multiple outcomes of interest which need to be measured at different scales. For example, the Watershared RCT in Bolivia which explored the impacts of payments for forest conservation on multiple outcomes, selected villages with their associated land as the randomization units (Wiik et al., 2019). However, this unit was inappropriate for estimating impacts of payments on water quality because watersheds did not align with community boundaries (Pynegar et al., 2018). Water quality outcomes could therefore spillover from upstream treated units to downstream control units, masking the effectiveness of the intervention (type iii spillover). To quantify this spillover would have required randomization of villages grouped within watershed, which would have been an unfeasibly large experiment.

Ensuring excludability

To use an RCT to estimate a causal effect without bias, a researcher must assume that the randomization procedure has no causal link to the outcome variable except through its effect on the treatment variable (i.e., the process of randomizing the treatment is not itself a potential source of confounding). That assumption is called the excludability assumption and a version of it must hold in any evaluation design, whether experimental or nonexperimental (Ferraro et al., 2019). When the assumption is violated, the control group no longer represents the counterfactual outcome for the treated group. Although excludability is satisfied in the idealized notion of an RCT, it may not be satisfied in real world RCTs (Kabeer, 2020; Kimmel et al., 2021).

In RCTs of behaviorally mediated conservation programs, a likely cause of a violation of the excludability assumption is behavioral effects which arise from participants knowing they are in a particular arm of an experiment (Pynegar et al., 2021). For example, in the KUL RCT, fishers in the control group, upon discovering that they were not selected to receive payments, may nevertheless seek to demonstrate that they can be conservation-oriented by releasing target species which they would not have done in the absence of the randomized payment design. In the Comoros RCT, landowners in the control group, upon discovering that they were not selected to receive payments, may reduce deforestation rates relative to their behavior in the absence of the RCT because they wish to show they are conservation-oriented, or they may instead clear more forest because they are upset about not being selected for the treatment group. To detect such behavioral effects, researchers can use qualitative research alongside the RCT (Table 2).

Excludability violations can also arise in RCTs when some units are lost to follow-up and so outcomes are not measured for all units (attrition), or when units do not comply with their treatment assignment (noncompliance). If the variables that affect attrition and noncompliance also affect outcomes, excludability is violated. In other words, these mechanisms can reintroduce confounders which RCT designs are meant to remove. To address attrition, researchers can try to directly control for the confounders or place bounds on the true impacts (Gerber & Green, 2012,Chapter 7; Table 2). To address noncompliance, researchers can analyze units "as assigned" rather than "as treated" to estimate an Intention-to-Treat effect (see Wiik et al 2019 for an example), or they can use randomization as an instrumental (surrogate) variable to estimate a Complier Average Casual Effect (Gerber & Green, 2012, Chapter 5-6), also known as a Local Average Treatment Effect.

Delivering external validity

A persistent critique of RCTs has been their limited external validity: the extent to which results can be generalized beyond the specific context in which the RCT was carried out (Barrett & Carter, 2010; Campbell, 1957)There is often a trade-off between how well the estimated effect would generalize to other times and places and how well an RCT can isolate the effect of an intervention (its internal validity; Krauss, 2021). For example, to allow the treatment to be randomized in a cost-effective manner, the researcher may end up selecting a nonrepresentative subsample of the target population. External validity can also be compromised when the values of the trial's outcome variables are affected by the process of being measured (the so-called Hawthorne effect in behavioral experiments; Pynegar et al., 2021).

External validity can be enhanced by running RCTs of the same intervention simultaneously in different locations to allow the influence of contextual variability to be explored. This practice has become more common in development economics over the past decade (Banerjee et al., 2015). A pioneering recent example from an area aligned to conservation (the impact of community monitoring on common pool resources: Ferraro & Agrawal, 2021) included six coordinated country-level studies and showed that such ambitious evaluations are possible in conservation. Similarly, the KUL RCT was run simultaneously in two different fishery contexts in two different places in Indonesia, allowing cross-context comparisons (Booth et al. IN REVIEW).

Less ambitious approaches to improving the external validity of an RCT exist. One example is to conduct a moderator analysis to explore the characteristics that drive heterogeneity in treatment effects between subgroups and use this understanding to infer how the results of an RCT might translate to a population with different distributions of those characteristics. For example, in an RCT exploring the impact of providing information to households on water use, owners responded more strongly than renters, suggesting the average effect in communities with a greater proportion of renters will be lower than observed in an RCT conducted in an area with a greater proportion of owners (Ferraro and Miranda, 2013). Similarly, mediator analysis sheds light on the mediating pathways through which the treatment effect arose in the RCT. For example, in an RCT of community monitoring on common pool resource use, the impact seemed to be mediated by information transmission to community management councils, suggesting that the treatment won't be as effective in communities that lack such councils (Ferraro and Agrawal, 2021).

2) Ethical and practical challenges

In addition to technical challenges, ethical and practical challenges can make it hard for conservation organizations to implement RCTs. These include the ethical challenges associated with randomization, as well as the practical challenges reconciling implementation of conservation with running an evaluation, and for conservation organizations to access appropriate technical expertise.

Ethical considerations

Where the intervention is likely to be beneficial, there are ethical concerns about allocating access to the intervention via randomization (Barrett & Carter, 2010). The ethical basis for running RCTs in medicine comes from the principle of *clinical equipoise*, which implies that RCTs are ethical if there is genuine uncertainty about whether a treatment is beneficial relative to the status quo treatment (Abramowicz & Szafarz, 2019). This may be difficult to satisfy in non-medical contexts where the intervention includes access to cash transfers, training, or other benefits, as is often the case in development, education, or in behaviorally mediated conservation programs. Similarly, conservationists may view it as unethical to withhold interventions which could conserve threatened species or habitats from some units for the sake of an experiment.

However, while the ethical challenges associated with randomization are real (Evans, 2023; Ravallion, 2020), it can also be argued that *not* randomizing when randomization is possible can also be unethical (Evans, 2023; Ferraro, et al., 2023). Passing up opportunities for learning via an RCT risks future misdirection of limited resources towards ineffective programs (Alpízar & Ferraro, 2020). Moreover, using conventional monitoring approaches with only observational data can lead to false conclusions. For example, in the KUL RCT Booth et al. (IN REVIEW) showed that a conventional monitoring approach, which quantified impacts based on numbers of sharks and wedgefish released, implied the program was hugely successful, with a 71% reduction in wedgefish mortality. However, the experimental data showed that payments also induced some vessels to increase their catches, eliminating the conservation benefits from the releases. In the absence of the RCT, the program staff would have observed the hundreds of live releases by participants and would have likely focused on scaling up the program. Instead, the program staff sought to adapt the program's design to eliminate its perverse incentives for vessels to increase their catch (Booth et al. IN REVIEW).

In practice, randomization can be a fair approach when an intervention cannot be applied to all units at the same time because of budgetary or logistical constraints (Evans, 2023). This justification is used in the Dahari RCT and the Mitsilo RCT, where there were insufficient resources to roll-out the interventions at larger scales. In such situations randomization may be preferred to allocation based on nepotistic or clientelist relationships (Asquith, 2020; Barrett, 2021). The risk of dissatisfaction among control groups can be further reduced by carefully explaining that robust evidence of the effectiveness of the intervention will make it easier to raise funds for wider rollout, by ensuring control groups are prioritized for implementation of the program after the RCT, by ensuring local scholars have leadership roles, and by maintaining good communication (Evans, 2023; Heard et al., 2017).

Where possible, a crossover or stepped wedge designs can avoid the ethical issues associated with randomization because everyone is exposed to both treatment and control conditions (Hemming et al., 2015). In the KUL RCT in Indonesian fisheries, a crossover design was adopted, where vessels were rotated between control and treatment arms every three months. This design was possible because the treatment - an incentive payment - could be switched on and off with low risk of spillovers between treated and control conditions, and the outcome variable - daily landings of target species - could be measured over short timeframes. A crossover design would be more challenging in RCTs with interventions that cannot be easily turned on and off (e.g., payments based on not clearing land such as in the Dahari RCT) or with outcomes that need to be measured over longer timeframes (e.g., avoided deforestation).

Reconciling conservation implementation and evaluation

A conservation organization's primary aim is to deliver conservation. When that aim comes into conflict with the needs of the RCT design, challenges arise.

Firstly, the conservation intervention cannot be implemented until the RCT is designed, which means that implementation may not be as quick as program staff desire. Although such delays happen in any prospective evaluation design, whether experimental or not, conservation practitioners may be unwilling to go through the necessary steps to set up an RCT because of the urgency of taking action to conserve biodiversity and because of funder pressures to start interventions immediately.

Secondly, implementing a program as an RCT can require changing the design of the intervention. In the Comoros RCT, Dahari considered augmenting the payments to farmers with community-level bonuses for local monitoring effort. However, the bonuses were dropped partly because they were not compatible with the RCT's randomization at the household level.

Thirdly, robust evaluation can pose a challenge for adaptive management – where changes to programmes are made via frequent informal tweaks as practitioners learn while doing (Salafsky et al., 2002). Given the slow pace of change for many ecological outcomes (Baylis et al., 2016), RCTs with ecological outcomes will likely need to run for at least a few years to detect change and it can be difficult for conservation practitioners to maintain the fidelity of a randomization design over several years. While the growing literature on 'adaptive designs' in clinical studies describes how interim assessments can be scheduled into RCT designs while maintaining the validity and integrity of the trial (Bhatt & Cyrus, 2016; Pallmann et al., 2018), such approaches require large numbers of randomization units.

Finally, randomization means conservationists must spread out their activities in space or time, which can make a program more difficult to implement. They can't, for example, start implementation in locations they have worked in before and know will be easy to work in, or start with easy to access locations and then slowly roll out the program to locations farther away.

Expertise and the challenges of transdisciplinary collaborations

Designing and implementing RCTs requires specialist skills not usually found in conservation organizations. Ferraro et al. (2023) suggested that trained experimentalists be fully embedded within implementing organizations to make the most of opportunities to design randomization into programmes. The benefits of embedding researchers in government conservation agencies for the generation and use of evidence has been discussed by others (Roux et al., 2019). However, it is unlikely that local, small to medium-sized implementing organizations will be able to maintain such specialist expertise in-house (Asquith, 2020; Sutherland, 2022; Vargas et al., 2022), and thus collaborations between conservation organizations and academic institutions will often be needed. While this collaboration can be beneficial, there can also be challenges because researchers and implementing organizations have different objectives and incentive structures (e.g. Jarvis et al., 2020).

3) Funding models and incentive structures

Prevailing funding models and incentive structures in conservation have not favored the use of RCTs (Alpízar & Ferraro, 2020; Asquith, 2020;Ferraro and Messer this issue).

Firstly, conducting an RCT can require substantial additional costs over and above implementation costs of the program (although such costs are often found in any high-quality evaluation design). Where baseline data collection is required, rather than relying on existing census or remotely sensed data, monitoring effort must be duplicated to include control units (however, baseline data collection is arguably even more important in non-randomized designs where such data is needed to allow the influence of confounders to be controlled for). Similarly, high-quality experimental design and subsequent analysis requires funding for people with appropriate skills (Ferraro, 2012), but again, expertise to analyze observational designs can be greater than that required to analyze RCTs.

Second, the timescales for well-designed RCTs may not match the typical timescales of conservation project funding. As noted in our summary of practical challenges, prospective evaluation designs, like an RCT, may require waiting years to measure the relevant outcomes. For example, the Watershared RCT in Bolivia, ran over five years (Pynegar et al., 2021; Wiik et al., 2019, 2020). Funders are often hesitant to fund the set-up of an RCT if their funding will not cover the endline. Only recently have some funders been willing to finance the entirety of large, expensive (>$1m), impact evaluations (e.g. (Ferraro & Agrawal, 2021). While this is encouraging to see, broader changes in funding structures are likely still necessary for large-scale randomized experiments to be widely used Ferraro, et al., 2023; Jones & Shreedhar, 2024).

Third, RCTs provide transparent, high-quality results about program impacts, but if those results do not imply program success, RCTs can create reputational risks for implementer organizations. This risk is exacerbated by the lack of a “safe to fail culture” in the conservation space (Catalano et al., 2018). Practitioners who are worried about the reputational risks from an evaluation that does not support the original theory of change may not wish to conduct an RCT precisely because RCT results are so simple to understand and hard to manipulate – in other words, RCTs leave fewer evaluator degrees of freedom for messaging. RCTs are further disincentivized because, although evidence of program effectiveness can help justify further investments in a program, few conservation donors require robust counterfactual designs, and some have failed to value them when organizations invest in them (Asquith, 2020). Greater use of RCTs in conservation will thus require a cultural shift whereby funders reward, rather than penalize, conservation organizations who invest in generating robust evidence and, as a result, are more modest in their claims (Jones & Shreedhar, 2024).

Steps needed for good design and reporting of RCTs

As in any high-quality evaluation, the interpretation of the results of an RCT depends on the quality of the design and the transparency of reporting (Miguel et al., 2014). In the previous sections, we introduced some of actions that are necessary to ensure high quality designs for RCTs that evaluate behaviorally mediated conservation programs. These actions are summarized in Figure 4. In this section, we consider additional actions that conservation programs can take to ensure that the results from RCTs are credible and transparent to stakeholders. These new actions are also in Figure 4.

To enhance the credibility of an RCT design, preregistration and pre-analysis plans are important. In medicine, it has long been standard practice that the existence and design of all RCTs should be preregistered on an official trials register. That practice arose as a result of scandals showing that negative findings from drug trials were being suppressed (Laine et al., 2007). Preregistration of study designs, ideally along with full analysis plans, not only reduce publication bias but also reduce the influence of hind-sight bias, which leads to practices such as hypothesizing after results are known (Munafò et al., 2017; Nosek et al., 2018). Thus, preregistration of RCTs, including defining outcomes, is increasingly standard in many fields . While there is no universally recognized registry for RCTs in conservation, pre-registration is certainly best practice. The three RCTs presented in this paper are pre-registered with the AEA RCT Registry (Dahari RCT) and AsPredicted (KUL RCT and Mitsilo RCT).

To further enhance credibility, conflicts of interest in conservation RCTs need to be dealt with carefully. In the medical field it is accepted that those involved in developing a treatment should have no influence over the design or implementation of RCTs testing those treatments (e.g. Dunn et al., 2016). However, such independence can be difficult to achieve in the context of RCTs of conservation programs where implementers will often need to be involved in the design and implementation of the RCT for logistical reasons (e.g. Asquith, 2020). Therefore, risk of conflicts of interest must be acknowledged and systems put in place to ensure the implementing organization cannot influence results for example by ensuring data collection is carried out by enumerators who are not employed by the implementing organization wherever possible (Eble et al., 2017).

In addition to taking actions to ensure high quality of RCT designs, RCT users also need to take actions to transparently report their results. In the medical literature, a CONSORT statement (Consolidated Standards of Reporting Trials; (Moher et al., 2010) accompanies all publications of RCT results. Although such statements are not standardized in the same way in other fields, they provide a useful checklist for the reporting of any RCT. For example, to highlight attrition or noncompliance with treatment assignment, it can be useful to include a flow diagram showing what happened to units throughout the implementation of the RCT.

Conclusions

Conservation programs operate in coupled human natural systems with interlinked drivers and feedbacks loops between social and environmental components. This complexity makes designing robust impact evaluation difficult. While the details in this paper might lead a reader to conclude that RCTs are beset by potential pitfalls, observational evaluation designs are even more technically challenging because of the difficulty in constructing valid counterfactuals of what would have happened without a conservation program. While there can be ethical challenges to randomizing an intervention believed to be beneficial (to threatened biodiversity or for local people), there are also ethical issues with not randomizing: scaling-up conservation programs without evaluating their impact risks wasting precious resources. We are not suggesting that RCTs are appropriate in all circumstances, but we firmly believe that greater use of randomization in conservation program evaluation would advance the effectiveness of conservation. There is also growing pressure for conservation to demonstrate a measurable impact, especially to access private sector sources of funding.

However, some of us have been calling for more RCTs of conservation programs for nearly 20 years (Ferraro & Pattanayak, 2006) and they remain rare. A huge body of literature and expertise from RCTs of other behaviorally mediated programs can help conservation as a field overcome the technical challenges to the design and implementation of RCTs. The barrier now is changes to incentive structures – particularly more pressure from funders to encourage conservation organizations to conduct high-quality evaluations of their approaches. We hope this paper will help convince both conservation organizations and funders that more use of RCTs to evaluate programs is needed, and that it will finally help to unlock the potential of this powerful evaluation approach in conservation.

Acknowledgement

This work was primarily funded by a grant from the Darwin Initiative Innovation Grant to Bangor University, the University of Oxford, and Dahari Comoros (DARNV004). Booth acknowledges Darwin grant 30-008. Ferraro acknowledges support from USDA's National Institute for Food and Agriculture (#2019-67023-29854). Rakotonarivo acknowledges funding from the European Union (Grant no. DCI-PANAF/2020/420-028), through the African Research Initiative for Scientific Excellence (ARISE) pilot programme implemented by the African Academy of Sciences with support from the European Commission and the African Union Commission. Jones, Doulton, Pynegar and Mohamad are all directly involved in the Dahari RCT, Booth and Ferraro are involved in the KUL RCT, and Rakotonarivo is involved in the Mitsilo RCT. Thanks to Owen Lewis, Nastazia Mohamadi and the Dahari forest team.

References

Abramowicz, M., & Szafarz, A. (2019). *Ethics of Randomized Controlled Trials: Should Economists Care about Equipoise?* https://doi.org/10.2139/ssrn.3465762

Adjognon, G. S., van Soest, D., & Guthoff, J. (2021). Reducing Hunger with Payments for Environmental Services (PES): Experimental Evidence from Burkina Faso. *American Journal of Agricultural Economics*, *103*(3), 831–857. https://doi.org/10.1111/AJAE.12150

Alpízar, F., & Ferraro, P. J. (2020). The environmental effects of poverty programs and the poverty effects of environmental programs: The missing RCTs. *World Development*, *127*, 104783. https://doi.org/10.1016/J.WORLDDEV.2019.104783

Arendt, F., & Matthes, J. (2016). Nature Documentaries, Connectedness to Nature, and Pro-environmental Behavior. *Environmental Communication*, *10*(4), 453–472. https://doi.org/10.1080/17524032.2014.993415

Asquith, N. (2020). Large-scale randomized control trials of incentive-based conservation: What have we learned? *World Development*, *127*, 104785. https://doi.org/10.1016/J.WORLDDEV.2019.104785

Baird, S., Bohren, A., McIntosh, C., & Ozler, B. (2017). Optimal Design of Experiments in the Presence of Interference*, Second Version. *PIER Working Paper Archive*. https://ideas.repec.org//p/pen/papers/16-025.html

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., Shapiro, J., Thuysbaert, B., & Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236), 1260799. https://doi.org/10.1126/science.1260799

Banerjee, A. V., Duflo, E., & Kremer, M. (2020). The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy. In K. Basu, D. Rosenblatt, & C. Sepúlveda (Eds.), *The State of Economics, the State of the World* (pp. 439–487). The MIT Press. https://direct.mit.edu/books/book/4917/chapter/624664/The-Influence-of-Randomized-Controlled-Trials-on

Barrett, C. B. (2021). On design-based empirical research and its interpretation and ethics in sustainability science. *Proceedings of the National Academy of Sciences*, *118*(29), e2023343118. https://doi.org/10.1073/pnas.2023343118

Barrett, C. B., & Carter, M. R. (2010). The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy*, *32*(4), 515–548. https://doi.org/10.1093/AEPP/PPQ023

Behaghel, L., Macours, K., & Subervie, J. (2019). How can randomised controlled trials help improve the design of the common agricultural policy? *European Review of Agricultural Economics*, *46*(3), 473–493. https://doi.org/10.1093/erae/jbz021

Bhatt, L. D., & Cyrus, M. (2016). Adaptive Designs for Clinical Trials. *New England Journal of Medicine*, *375*(1), 65–74. https://doi.org/10.1056/NEJMra1510061

Buntaine, M. T., Zhang, B., & Hunnicutt, P. (2021). Citizen monitoring of waterways decreases pollution in China by supporting government action and oversight. *Proceedings of the National Academy of Sciences*, *118*(29), e2015175118. https://doi.org/10.1073/pnas.2015175118

Byerly, H., Balmford, A., Ferraro, P. J., Hammond Wagner, C., Palchak, E., Polasky, S., Ricketts, T. H., Schwartz, A. J., & Fisher, B. (2018). Nudging pro-environmental behavior: evidence and opportunities. *Frontiers in Ecology and the Environment*, *16*(3), 159–168. https://doi.org/10.1002/fee.1777

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

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. https://doi.org/10.1111/cobi.13045

Chaves, W. A., Valle, D. R., Monroe, M. C., Wilkie, D. S., Sieving, K. E., & Sadowsky, B. (2018). Changing Wild Meat Consumption: An Experiment in the Central Amazon, Brazil. *Conservation Letters*, *11*(2), e12391. https://doi.org/10.1111/conl.12391

Cheng, S. H., McKinnon, M. C., Masuda, Y. J., Garside, R., Jones, K. W., Miller, D. C., Pullin, A. S., Sutherland, W. J., Augustin, C., Gill, D. A., Wongbusarakum, S., & Wilkie, D. (2020). Strengthen causal models for better conservation outcomes for human well-being. *PLOS ONE*, *15*(3), e0230495. https://doi.org/10.1371/journal.pone.0230495

Christie, A. P., Abecasis, D., Adjeroud, M., Alonso, J. C., Amano, T., Anton, A., Baldigo, B. P., Barrientos, R., Bicknell, J. E., Buhl, D. A., Cebrian, J., Ceia, R. S., Cibils-Martina, L., Clarke, S., Claudet, J., Craig, M. D., Davoult, D., De Backer, A., Donovan, M. K., … Sutherland, W. J. (2020). Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. *Nature Communications*, *11*(1), 6377. https://doi.org/10.1038/s41467-020-20142-y

Connolly, P., Biggart, A., Miller, Dr. S., & O'Hare, L. (2017). *Using Randomised Controlled Trials in Education*. https://sk.sagepub.com/books/using-randomised-controlled-trials-in-education

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

Demarchi, G., Subervie, J., Carrilho, C., Catry, T., Pfefer, A., & Delacote, P. (2024). Greater Flexibility in Payments for Ecosystem Services: Evidence from an RCT in the Amazon. *Hal Open Science*. https://hal.inrae.fr/hal-04462983

Donner, A., & Klar, N. (2004). Pitfalls of and Controversies in Cluster Randomization Trials. *American Journal of Public Health*, *94*(3), 416–422. https://doi.org/10.2105/AJPH.94.3.416

Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 Using Randomization in Development Economics Research: A Toolkit. *Handbook of Development Economics*, *4*, 3895–3962. https://doi.org/10.1016/S1573-4471(07)04061-2

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. https://doi.org/10.1111/csp2.280

Eble, A., Boone, P., & Elbourne, D. (2017). On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics. *The World Bank Economic Review*, *31*(3), 687–707. https://doi.org/10.1093/wber/lhw034

Evans, D. K. (2023). Towards improved and more transparent ethics in randomised controlled trials in development social science. *Journal of Development Effectiveness*, *0*(0), 1–11. https://doi.org/10.1080/19439342.2023.2196978

Ferraro, P. J. (2009). Counterfactual thinking and impact evaluation in environmental policy. *New Directions for Evaluation*, *2009*(122), 75–84. https://doi.org/10.1002/ev.297

Ferraro, P. J. (2012). *Designing projects to create evidence and catalyze investments to secure global environmental benefits*. https://stapgef.org/sites/default/files/stap/wp-content/uploads/2013/05/Experimental-Design.pdf

Ferraro, P. J., & Agrawal, A. (2021). Synthesizing evidence in sustainability science through harmonized experiments: Community monitoring in common pool resources. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(29), e2106489118. https://doi.org/10.1073/PNAS.2106489118/ASSET/0DACE039-2A75-4D0E-BB9F-B3EC78F706D0/ASSETS/IMAGES/LARGE/PNAS.2106489118FIG01.JPG

Ferraro, P. J., Cherry, T. L., Shogren, J. F., Vossler, C. A., Cason, T. N., Flint, H. B., Hochard, J. P., Johansson-Stenman, O., Martinsson, P., Murphy, J. J., Newbold, S. C., Thunström, L., van Soest, D., van 't Veld, K., Dannenberg, A., Loewenstein, G. F., & van Boven, L. (2023). Create a culture of experiments in environmental programs. *Science*, *381*(6659), 735–737. https://doi.org/10.1126/science.adf7774

Ferraro, P. J., Cherry, T. L., Shogren, J. F., Vossler, C. A., Cason, T. N., Flint, H. B., Hochard, J. P., Johansson-Stenman, O., Martinsson, P., Murphy, J. J., Newbold, S. C., Thunström, L., van Soest, D., van't Veld, K., Dannenberg, A., Loewenstein, G. F., & van Boven, L. (2023). Create a culture of experiments in environmental programs. *Science*, *381*(6659), 735–737. https://doi.org/10.1126/SCIENCE.ADF7774

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*(1), 495–517. https://doi.org/10.1146/annurev-environ-101813-013230

Ferraro, P. J., & Pattanayak, S. K. (2006). Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments. *PLOS Biology*, *4*(4), e105. https://doi.org/10.1371/journal.pbio.0040105

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. https://doi.org/10.1073/PNAS.1805563115/SUPPL_FILE/PNAS.1805563115.SAPP.PDF

Gerber, A. S., & Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation. Chapter*. https://www.amazon.com/Field-Experiments-Design-Analysis-Interpretation/dp/0393979954

Gertler, P., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact evaluation in practice* (2nd ed.). Inter-American Development Bank and World Bank.

Glennerster, R., & Takavarasha, K. (2013). Running Randomized Evaluations: A Practical Guide. In *Running Randomized Evaluations*. Princeton University Press. https://www.degruyter.com/document/doi/10.1515/9781400848447/html

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. https://doi.org/10.1016/j.ecolecon.2019.106404

Heard, K., O'Toole, E., Naimpally, R., & Bressler, L. (2017). Real-world challenges to randomization and their solutions. In *J-PAL*. https://www.povertyactionlab.org/sites/default/files/research-resources/2017.04.14-Real-World-Challenges-to-Randomization-and-Their-Solutions.pdf

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, *350*, h391. https://doi.org/10.1136/bmj.h391

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. https://doi.org/10.1126/science.aan0568

Jayachandran, S., de Laat, J., Lambin, E., & Stanton, C. (2016). *Cash for Carbon: A Randomized Controlled Trial of Payments for Ecosystem Services to Reduce Deforestation*. https://doi.org/10.3386/w22378

Jones, D. S., & Podolsky, S. H. (2015). The history and fate of the gold standard. *The Lancet*, *385*(9977), 1502–1503. https://doi.org/10.1016/S0140-6736(15)60742-5

Jones, J. P. G., & Shreedhar, G. (2024). The causal revolution in biodiversity conservation. *Nature Human Behaviour 2024*, 1–4. https://doi.org/10.1038/s41562-024-01897-6

Kabeer, N. (2020). 'Misbehaving' RCTs: The confounding problem of human agency. *World Development*, *127*, 104809. https://doi.org/10.1016/j.worlddev.2019.104809

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. https://doi.org/10.1016/j.tree.2021.08.008

Krauss, A. (2021). Assessing the Overall Validity of Randomised Controlled Trials. *International Studies in the Philosophy of Science*, *34*(3), 159–182. https://doi.org/10.1080/02698595.2021.2002676

Laine, C., Horton, R., DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Godlee, F., Haug, C., Hébert, P. C., Kotzin, S., Marusic, A., Sahni, P., Schroeder, T. V., Sox, H. C., Van Der Weyden, M. B., & Verheugt, F. W. (2007). Clinical trial registration: looking back and moving ahead. *Lancet*, *369*(9577), 1909–1911. https://doi.org/10.1016/S0140-6736(07)60894-0

Ma, Z., Bauchet, J., Steele, D., Godoy, R., Radel, C., & Zanotti, L. (2017). Comparison of Direct Transfers for Human Capital Development and Environmental Conservation. *World Development*, *99*, 498–517. https://doi.org/10.1016/j.worlddev.2017.05.030

Matthews, J. N. S. (2006). *Introduction to Randomized Controlled Clinical Trials* (2nd ed.). Chapman and Hall/CRC.

Meyfroidt, P., Börner, J., Garrett, R., Gardner, T., Godar, J., Kis-Katos, K., Soares-Filho, B. S., & Wunder, S. (2020). Focus on leakage and spillovers: informing land-use governance in a tele-coupled world. *Environmental Research Letters*, *15*(9), 090202. https://doi.org/10.1088/1748-9326/ab7397

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting Transparency in Social Science Research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, c869. https://doi.org/10.1136/bmj.c869

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. In *Nature Human Behaviour* (Vol. 1, Issue 1, pp. 1–9). Nature Publishing Group. https://doi.org/10.1038/s41562-016-0021

Nielsen, K. S., Marteau, T. M., Bauer, J. M., Bradbury, R. B., Broad, S., Burgess, G., Burgman, M., Byerly, H., Clayton, S., & Espelosin, D. (2021). Biodiversity conservation as a promising frontier for behavioural science. *Nature Human Behaviour*, 1–7.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Ockendon, N., Amano, T., Cadotte, M., Downey, H., Hancock, M. H., Thornton, A., Tinsley-Marshall, P., & Sutherland, W. J. (2021). Effectively integrating experiments into conservation practice. *Ecological Solutions and Evidence*, *2*(2), e12069. https://doi.org/10.1002/2688-8319.12069

Oyanedel, R., Gelcich, S., & Milner-Gulland, E. j. (2020). Motivations for (non-)compliance with conservation rules by small-scale resource users. *Conservation Letters*, *13*(5), e12725. https://doi.org/10.1111/conl.12725

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V, Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C., & Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, *16*(1), 29. https://doi.org/10.1186/s12916-018-1017-7

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*, *2018*(10). https://doi.org/10.7717/peerj.5753

Ravallion, M. (2020). Should the randomistas(continue to) rule? In *Randomized Control Trials in the Field of Development: A Critical Perspective* (pp. 47–78). Oxford University Press. https://doi.org/10.1093/OSO/9780198865360.003.0003

Roux, D. J., Kingsford, R. T., Cook, C. N., Carruthers, J., Dickson, K., & Hockings, M. (2019). The case for embedding researchers in conservation agencies. *Conservation Biology*, *33*(6), 1266–1274. https://doi.org/10.1111/cobi.13324

Salafsky, N., Margoluis, R., Redford, K. H., & Robinson, J. G. (2002). Improving the Practice of Conservation: a Conceptual Framework and Research Agenda for Conservation Science. *Conservation Biology*, *16*(6), 1469–1479. https://doi.org/10.1046/j.1523-1739.2002.01232.x

Sasaki, S., Kubo, T., & Kitano, S. (2024). Prosocial and Financial Incentives for Biodiversity Conservation: A Field Experiment Using a Smartphone App. *SSRN*. https://doi.org/10.2139/SSRN.4756100

Smith, R., Ockendon, N., & Sutherland, W. (2023). Conservation practitioners frequently do test the effectiveness of their actions, but this needs to become routine . *Science*, *eLetter*. https://doi.org/doi/10.1126/science.adf7774#elettersSection

Styles, B., & Torgerson, C. (2018). Randomised controlled trials (RCTs) in education research – methodological debates, questions, challenges. *Educational Research*. https://www.tandfonline.com/doi/abs/10.1080/00131881.2018.1500194

Sutherland, W. J. (2022). *Transforming Conservation: A Practical Guide to Evidence and Decision Making*. Open Book Publishers. https://www.openbookpublishers.com/books/10.11647/obp.0321

Sutherland, W. J., & Wordley, C. F. R. (2018). A fresh approach to evidence synthesis. *Nature 2021 558:7710*, *558*(7710), 364–366. https://doi.org/10.1038/d41586-018-05472-8

Vargas, M. T., Garcia, M., Vidaurre, T., Carrasco, A., Araujo, N., Medema, C., Asquith, N., Pynegar, E., Tobon, C., Manco, Y., Ma, Z., Bauchet, J., Grillos, T., & McWherter, B. (2022). The researcher–practitioner symbiosis: Evolving mutualisms from parachutes. *Conservation Science and Practice*, *4*(5), e596. https://doi.org/10.1111/csp2.596

Vorlaufer, T., Engel, S., de Laat, J., & Vollan, B. (2023). Payments for ecosystem services did not crowd out pro-environmental behavior: Long-term experimental evidence from Uganda. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(18). https://doi.org/10.1073/PNAS.2215465120

Wardropper, C. B., Esman, L. A., Harden, S. C., Masuda, Y. J., Ranjan, P., Weigel, C., Ferraro, P. J., Prokopy, L. S., & Reddy, S. M. W. (2022). Applying a “fail-fast” approach to conservation in US agriculture. *Conservation Science and Practice*, *4*(3), e619. https://doi.org/10.1111/CSP2.619

Webber, S., & Prouse, C. (2018). The New Gold Standard: The Rise of Randomized Control Trials and Experimental Development. *Economic Geography*, *94*(2), 166–187. https://doi.org/10.1080/00130095.2017.1392235

Weigel, C., Harden, S., Masuda, Y. J., Ranjan, P., Wardropper, C. B., Ferraro, P. J., Prokopy, L., & Reddy, S. (2021). Using a randomized controlled trial to develop conservation strategies on rented farmlands. *Conservation Letters*, *14*(4), e12803. https://doi.org/10.1111/conl.12803

Wiik, E., d'Annunzio, R., Pynegar, E., Crespo, D., Asquith, N., & Jones, J. P. G. (2019). Experimental evaluation of the impact of a payment for environmental services program on deforestation. *Conservation Science and Practice*, e8. https://doi.org/10.1002/csp2.8

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. *Conservation Biology*, *34*(5), 1076–1088. https://doi.org/10.1111/cobi.13508

Wilebore, B., & Coomes, D. (2016). Combining spatial data with survey data improves predictions of boundaries between settlements. *Applied Geography*, *77*, 1–7. https://doi.org/10.1016/j.apgeog.2016.09.007

Wilebore, B., Voors, M., Bulte, E., Coomes, D., & Kontoleon, A. (2019). Unconditional Transfers and Tropical Forest Conservation. Evidence from a Randomized Control Trial in Sierra Leone. *American Journal of Agricultural Economics*.

Wren-Lewis, L., Becerra-Valbuena, L., & Houngbedji, K. (2020). Formalizing land rights can reduce forest loss: Experimental evidence from Benin. *Science Advances*, *6*(26), eabb6914. https://doi.org/10.1126/sciadv.abb6914

Table 1: Typology of RCT evaluations in conservation classified according to whether an intervention directly impacts biodiversity or is behaviorally mediated, and whether measured outcomes are primarily ecological or social. RCTs concerning interventions with a direct impact on biodiversity (top row) are the mainstay of applied ecology and are relatively well understood (though see Kimmel et al., 2021). This paper therefore focuses on RCTs of behaviorally mediated programs (the bottom row).

		Outcomes	
		Ecological (e.g. habitat extent, condition, species trends, vital rates of populations, provisioning or regulating ecosystem services)	**Social** (e.g. attitude, behavior, wellbeing, multidimensional poverty, cultural ecosystem services)
Intervention	**Direct**	Christie et al. (2020) identify 736 RCTs in the *Conservation Evidence* database, Smith et al., 2023 suggested nearer 1600 are now included.	We have found no examples of RCTs in this category.
	Behaviorally mediated	• Impact of paying farmers in Uganda not to deforest on forest loss (Jayachandran et al., 2017). • Impact of incentivizing farmers in Bolivia not to deforest and to keep cattle out of riparian forest on water quality (Pynegar et al., 2018) and deforestation (Wiik et al., 2019). • Impact of unconditional cash transfers to farmers in Siera Leone on deforestation (Wilebore et al., 2019). • Impact of granting formal land tenure to farmers in Benin on deforestation rate (Wren-Lewis et al., 2020). • Impact of paying landowners in Brazil not to deforest, with differing levels of strictness associated with agreements, on deforestation rate (Demarchi et al., 2024). • Impact of enlisting volunteers to monitor water quality on pollution levels in Chinese waterways (Buntaine et al., 2021) • Impact of paying farmers to sow cover crops on farmland in the central United States (Weigel et al., 2021).	• Impact of social marketing and discount coupons on replacing wild meat consumption with chicken (Chaves et al., 2018). • Impact of watching a documentary on attitudes to nature (Arendt & Matthes, 2016). • Impact of watching a documentary on behaviors relating to marine pollution (Dunn et al., 2020). • Impact of incentivizing farmers in Bolivia not to deforest and to keep cattle out of riparian forest on environmental values(Grillos et al., 2019). • Long-term impact of a payments for ecosystem services programme on pro-environmental behavior, beliefs and motivations (Vorlaufer et al., 2023). • Impact of cash transfers to reforest in Burkina Faso on social welfare (Adjognon et al., 2021). • Impact of incentivizing contribution to citizen science through incentivization on using a smartphone app (Sasaki et al., 2024).

Table 2: Technical challenges with the design of RCTs of conservation programs and potential solutions.

Topic	**Specific Issue**	**Solution**
Selection of randomization unit.	The unit which is appropriate for measuring outcomes is large, meaning few units are available.	Select a proxy outcome which can be measured with a smaller unit, or do not try to evaluate that outcome.
	Different outcomes have different appropriate units.	Fully cluster units at which some outcomes are measured within units at which other outcomes are measured.
	Power analysis shows that the minimal detectable effect is too large to be useful (i.e. the experiment is underpowered).	Stratify on baseline characteristics likely to be correlated with the outcome to increase power. Coordinate with other programs where possible to boost power (and improve external validity). However, if the number of units is low, an RCT may not be the appropriate evaluation design.
Avoiding interference: i.e. ensuring the outcome of a unit depends only on its own treatment status, not on the status of other units.	Spillover between treatment and control units will bias the measured treatment effect.	Conduct qualitative research to better understand the social context and then select units which will minimize such spillovers. Consider a cluster RCTS, where some units within treatment clusters are left untreated to quantify spillovers. If this isn't possible, select intermediate outcomes from the theory of change to measure instead.
Ensuring excludability: i.e. that factors influencing treatment assignment have no causal link to the outcome except through their effects on treatment.	Participants behave differently because they know what arms of the experiment they are in (i.e. behavioral effects confound the RCT).	Conduct qualitative research alongside the RCT to provide insights into the extent to which this has occurred.
	Some units are lost to follow-up.	Control for the characteristics correlated with attrition, or use understanding of which units have been lost to follow-up to place bounds on the true impacts.
	Not all units receive the treatment they should have received according to randomization (there is noncompliance with treatment assignment).	Analyze units "as assigned" rather than "as treated" to estimate an Intention-to-Treat effect, or use randomization as an instrumental (surrogate) variable to estimate a Complier Average Casual Effect, also known as a Local Average Treatment Effect.
Ensuring the results can be generalized beyond the specific context in which the RCT was carried out.	Participants in the RCT behave differently because of the experiment (behavioral effects reduce external validity).	Conduct qualitative research alongside the RCT to provide insights into the extent to which this has occurred.
	The RCT is a major investment in a single context which tells you little about many other contexts where the intervention could be applied.	Conduct coordinated RCTs in multiple contexts concurrently. OR (simpler), explore the characteristics that drive heterogeneity in treatment effects between subgroups and use this understanding to infer how the results of an RCT might translate to different contexts. And/or explore mediating pathways through which the treatment effect arose in the RCT.

Population: Farmers with land on the forest frontier.

Intervention: Payment (though local micro-credit association) to individual farmers conditional on land management actions.

Comparator: Payment not offered to control units.

Outcomes: Forest cover, regeneration of native vegetation, household wellbeing, attitudes to conservation.

The Dahari RCT: The Comorian NGO Dahari has developed a conditional cash transfer programme for landowners who own forested land and who sign agreements not to deforest or cut trees as an RCT. The randomization unit is the farmer together with their eligible parcels, outcomes are measured at the scale of farmers and eligible parcels.

Population: Small-scale fishers which frequently capture endangered species.

Intervention: Fishers are offered cash payments for safely releasing Critically Endangered hammerhead sharks and wedgefish.

Comparator: Payment not offered (treatment and control rotated every 6 months).

Outcomes: Retained catches of sharks and wedgefish (a proxy for mortality), attitudes and behavioral beliefs, subjective wellbeing.

The Kebersamaan Untuk Lautan (KUL) RCT: An Indonesian NGO, KUL, is assessing the impact of their compensate-to-release scheme on marine biodiversity and wellbeing outcomes in small-scale fisheries in Indonesia. The randomization unit is the vessel. Catches are measured at the scale of the vessel and attitudes, beliefs, and wellbeing at the scale of individual fishers.

Population: Members of Village Savings and Loan Associations (VSLAs) in forest frontier areas.

Intervention: Credit injection and Farmer Business School Training (improved-farming techniques and business skills) offered via VSLAs.

Comparator: One group gets the credit injection, one gets the credit injection and training, the control group get neither.

Outcomes: income, various indicators of welfare (assets, multidimensional poverty index, food security) self-reported farming practices, burning of farmland.

The Mitsilo RCT: The Malagasy research lab Mitsilo is running an RCT to explore the impacts of providing support through Villages Savings and Loans Associations on well-being, agricultural decisions, and the use of fire among forest-edge communities in Madagascar. The randomization unit is the village with one or more VSLA clustered within it. Socio-economic outcomes are measured at the scale of the individual, while burning is measured at the scale of the zone of influence of the village.

Figure 1. Examples of ongoing RCTs of behaviorally mediated conservation programs: a) The Dahari RCT implemented by NGO Dahari in collaboration with Bangor University and the University of Oxford, b) The KUL RCT implemented by Yayasan Kebersamaan Untuk Lautan with evaluation led by University of Oxford and IPB University in Indonesia c) The Mitsilo RCT implemented by GIZ with the University of Antananarivo in Madagascar. The causal question each aims to answer is presented in the PICO (Population, Intervention, Comparator, Outcomes) framework.

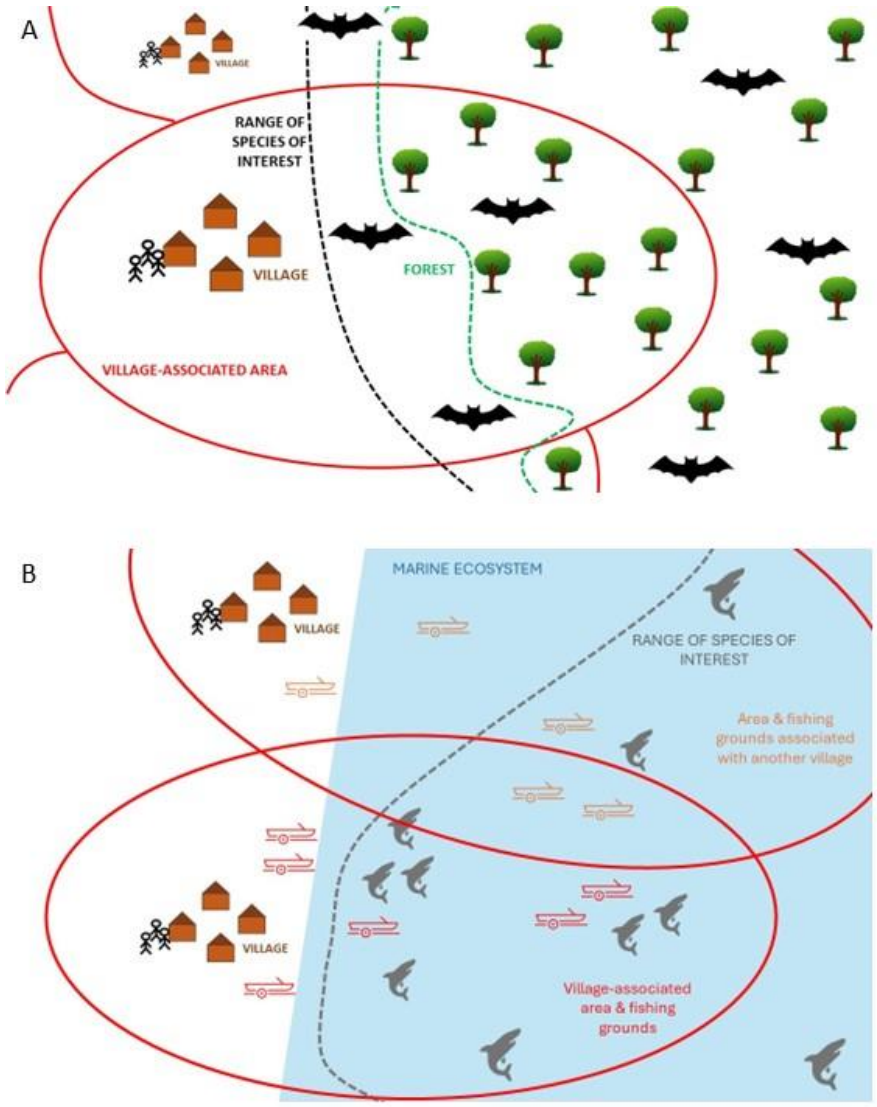
A
VILLAGE
RANGE OF SPECIES OF INTEREST
VILLAGE
FOREST
VILLAGE-ASSOCIATED AREA
B
MARINE ECOSYSTEM
VILLAGE
RANGE OF SPECIES OF INTEREST
Area & fishing grounds associated with another village
VILLAGE
Village-associated area & fishing grounds

Figure 2. Selecting the randomization unit in conservation RCTs is challenging because the unit at which it is appropriate to implement an intervention may differ from the unit at which it is appropriate to measure outcomes. We illustrate the challenge using the context of the Dahari RCT in the Comoros (A) and the KUL RCT in Indonesia (B). Outcomes such as human wellbeing or attitudes can be measured with individuals clustered within villages (A and B). To measure the impact of the intervention on conservation outcomes requires that there is a defined area primarily influenced by those living in a specific randomization unit such as a village, and that outcomes can be measured at that scale. In the Dahari RCT, a defined area might be possible for deforestation outcomes, but would not be possible for outcomes of wide-ranging target species, such as the critically endangered Livingstone fruit bat (A). Ecological outcomes cannot be measured for the KUL RCT directly because target species are wide-ranging and fishing grounds are dynamic (B).

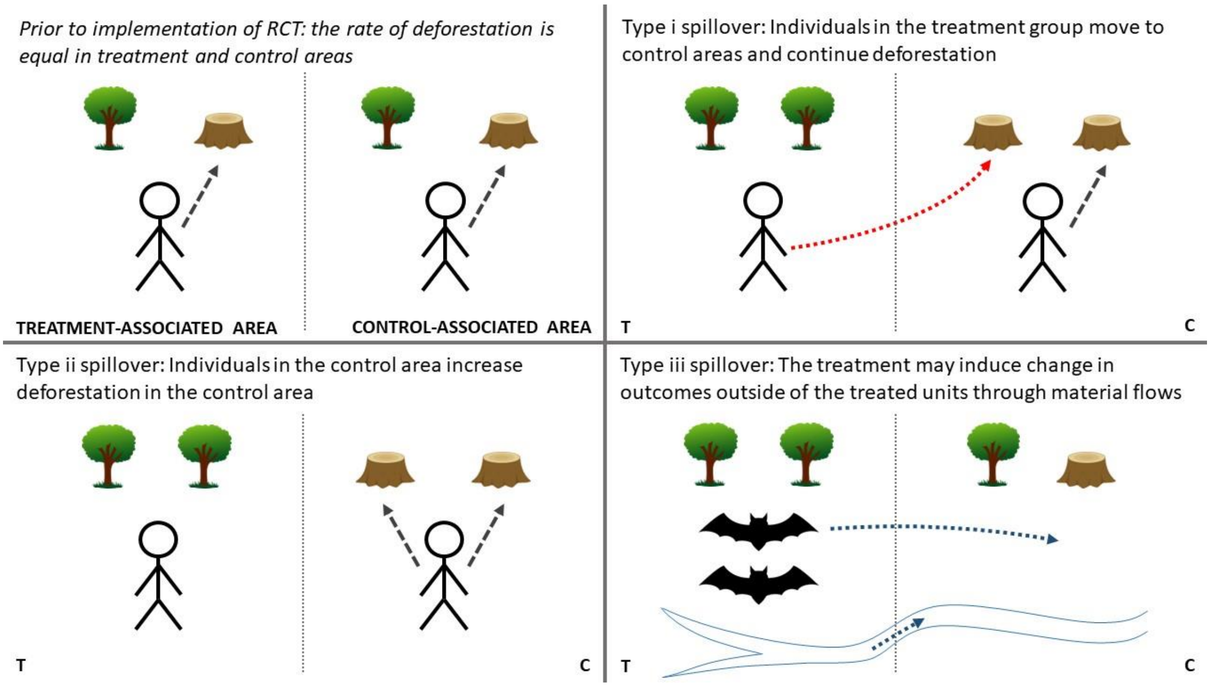
Prior to implementation of RCT: the rate of deforestation is equal in treatment and control areas
TREATMENT-ASSOCIATED AREA
CONTROL-ASSOCIATED AREA
Type i spillover: Individuals in the treatment group move to control areas and continue deforestation
T
C
Type ii spillover: Individuals in the control area increase deforestation in the control area
T
C
Type iii spillover: The treatment may induce change in outcomes outside of the treated units through material flows
T
C

Figure 3. Three distinct mechanisms may generate spillovers in behaviorally mediated conservation RCTs, violating the assumption of no interference among units. We illustrate this using the context of the Dahari RCT to evaluate the impact of conditional cash transfer program on deforestation rates and associated outcomes. Type i spillovers: people are displaced from treatment to control areas where they continue to clear land. Type ii spillovers: people in the control area increase their rate of land clearance to meet unmet demand from the treatment area. Type iii spillovers: outcomes spillover from treatment to control. Type i and ii spillovers will tend to result in overestimates of the treatment effect while Type iii spillovers will tend to result in underestimates.

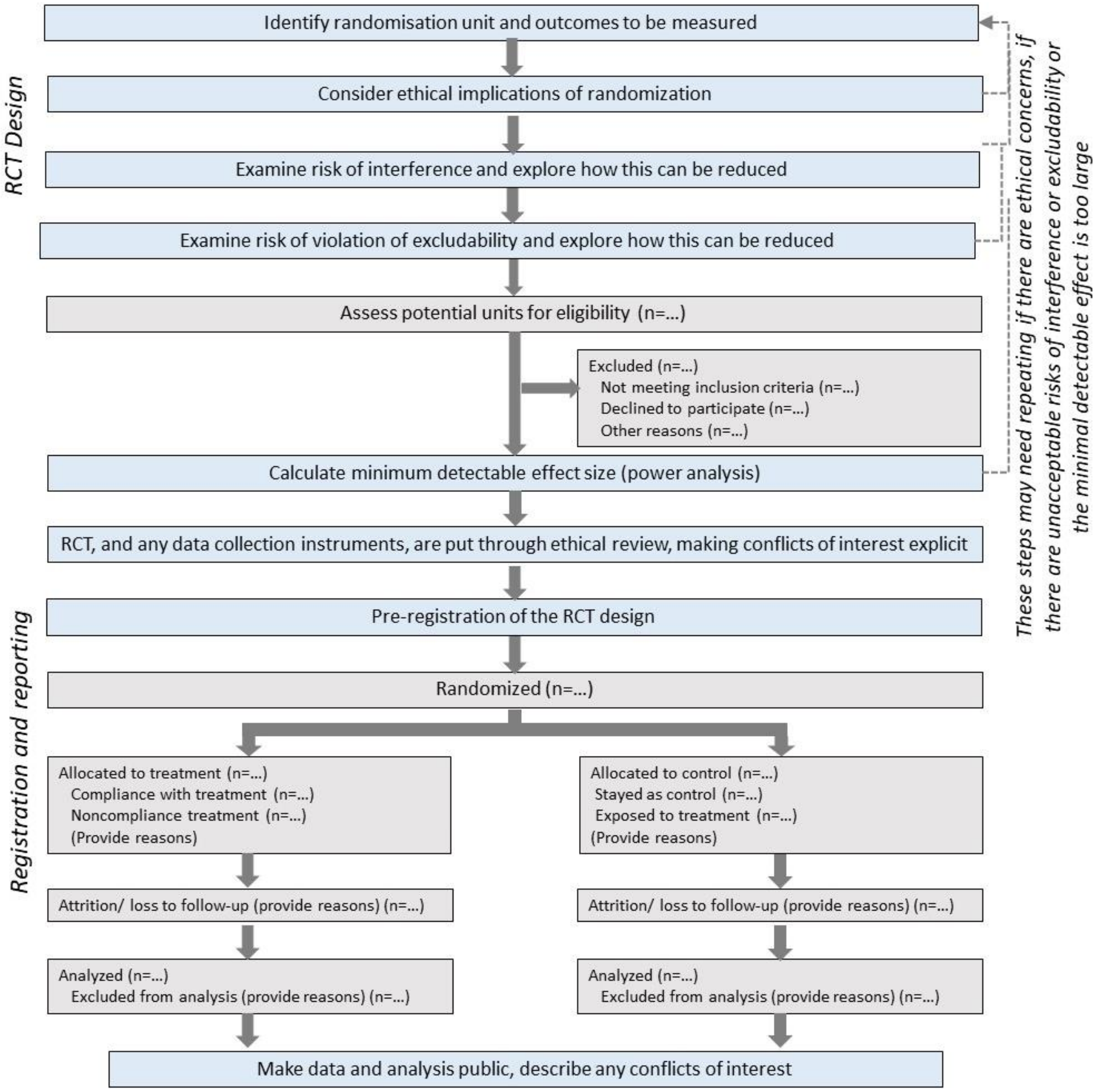
RCT Design
Identify randomisation unit and outcomes to be measured
Consider ethical implications of randomization
Examine risk of interference and explore how this can be reduced
Examine risk of violation of excludability and explore how this can be reduced
Assess potential units for eligibility (n=...)
Excluded (n=...)
Not meeting inclusion criteria (n=...)
Declined to participate (n=...)
Other reasons (n=...)
Calculate minimum detectable effect size (power analysis)
RCT, and any data collection instruments, are put through ethical review, making conflicts of interest explicit
These steps may need repeating if there are ethical concerns, if there are unacceptable risks of interference or excludability or the minimal detectable effect is too large
Registration and reporting
Pre-registration of the RCT design
Randomized (n=...)
Allocated to treatment (n=...)
Compliance with treatment (n=...)
Noncompliance treatment (n=...)
(Provide reasons)
Allocated to control (n=...)
Stayed as control (n=...)
Exposed to treatment (n=...)
(Provide reasons)
Attrition/ loss to follow-up (provide reasons) (n=...)
Attrition/ loss to follow-up (provide reasons) (n=...)
Analyzed (n=...)
Excluded from analysis (provide reasons) (n=...)
Analyzed (n=...)
Excluded from analysis (provide reasons) (n=...)
Make data and analysis public, describe any conflicts of interest

Figure 4. Flow diagram showing the steps which those designing an RCT of a behaviorally mediated conservation program need to go through, highlighting the importance of pre-registration and transparent reporting. The content of the standard CONSORT flow diagram (Moher et al., 2010) is shown in the grey boxes. The dotted lines represent what is often an iterative process.