What role should Randomised Control Trials play in providing the evidence base underpinning conservation?

Author List: Edwin L. Pynegar[a], James M. Gibbons[a], Nigel M. Asquith[b, c], Julia P. G. Jones[a]

Affiliations and Addresses:

[a]School of Environment, Natural Resources and Geography, Bangor University, Bangor, Gwynedd LL57 2UW, UK

[b]Harvard Forest, 324 N Main St, Petersham, MA 01366, USA

[c]Sustainability Science Program, Harvard Kennedy School, 79 John F. Kennedy St, Box 81, Cambridge, MA 02138, USA.

Corresponding Author: Edwin L. Pynegar

Postal address: School of Environment, Natural Resources and Geography, Bangor University, Deiniol Road, Bangor, Gwynedd LL57 2UW, UK

Email: edwin.pynegar@gmail.com

Abstract

There is general agreement that conservation decision-making should be evidence-informed, but many evaluations of intervention effectiveness do not attempt to account for confounding variables and so provide weak evidence. Randomised Control Trials (RCTs), in which experimental units are randomly allocated to treatment or control groups, offer an intuitive means of calculating the effect size of an intervention through establishing a reliable counterfactual and avoid the pitfalls of alternative quasi-experimental approaches. However, RCTs may not be the most appropriate way to answer some kinds of evaluation question, are not feasible in all circumstances, and factors such as spillover and behavioural effects risk prejudicing their quality. Some of these challenges may be greater in situations where the intervention aims to influence ecological outcomes through changing human behaviour (socio-ecological interventions). The external validity – the extent to which findings are generalizable – of RCT impact evaluation has also been questioned. We offer guidance and a series of criteria for deciding when RCTs may be a useful approach for evaluating the impact of conservation interventions, and what must be considered to ensure an RCT is of high quality. We illustrate this with examples from one of the few RCTs of a socio-ecological intervention – an incentive-based conservation program in the Bolivian Andes. Those who care about evidence-informed environmental management should aim to avoid a re-run of the polarized debate surrounding RCTs' use in fields such as development economics and take a pragmatic approach to impact evaluation, while also actively integrating learning from these fields. If this can be achieved, they will have a useful role to play in robust impact evaluation.

Introduction

Land managers, policymakers and other stakeholders make decisions about how ecosystems should be managed. There are increasing calls that such decisions should be firmly rooted in robust evidence (Sutherland et al., 2004; Segan et al., 2011; Baylis et al., 2016). Reasons why current decisions may not be evidence-based include decision makers' lack of access to evidence (Pullin et al., 2004) and inertia to changing established practices (Sutherland et al., 2004). However there are also clear limitations in the available evidence on the likely impacts of potential conservation interventions in a given situation (Ferraro & Pattanayak, 2006; Pattanayak, Wunder & Ferraro, 2010).

Impact evaluation (described by the World Bank as assessment of changes in outcomes of interest attributable to specific interventions; Independent Evaluation Group 2012) requires a counterfactual: an understanding of what would have occurred without that intervention (Margoluis et al., 2009; Miteva, Pattanayak & Ferraro, 2012; Ferraro & Hanauer, 2014; Baylis et al., 2016). It is well recognized that simple before-and-after comparison of units exposed to the intervention is flawed, as some factor other than the intervention may have caused the change in the outcome of interest (Ferraro & Hanauer, 2014; Baylis et al., 2016). Comparing groups exposed and not exposed to the intervention is also flawed as the groups may differ in other, potentially unobserved, ways that affect the outcome.

One solution is to replace simple post-project monitoring with more robust quasi-experiments, in which a variety of approaches may be used to construct a counterfactual scenario statistically. *Statistical matching*, including *propensity score matching*, involves comparing outcomes in units where an intervention is implemented with outcomes in similar (statistically selected) units lacking the intervention. This is increasingly used for conservation impact evaluations such as determining the effectiveness of a sustainable agriculture program (Margoluis et al., 2001) and in investigating the impact of national park establishment (Andam et al., 2008) or Community Forest Management (Rasolofoson et al., 2015) on deforestation. Other quasi-experimental approaches include *instrumental variables* (where easily observable variables correlated with the intervention but not the outcome are used as a proxy for the treatment), the *regression-discontinuity* approach (which compares outcomes of interest in units just above and below an initial eligibility criterion for implementation of the intervention:; as the criterion is arbitrary, units on either side will be essentially identical other than in implementation of the intervention), and *difference-in-differences* (which compares changes in outcomes in units exposed to an intervention with changes in a comparison group which was not exposed). Butsic *et al.* (2017) provide much more information on quasi-experiments' use in a conservation context.

Quasi-experiments should, and increasingly do, have a major role to play in conservation impact evaluation, and in some situations will be the only robust option available to evaluators. Their use has become substantially more common in recent years, which should be greatly welcomed, and meta-analyses of the effectiveness of certain interventions have recently begun to be published based upon quasi-experimental analyses (Samii et al., 2014; also see Börner et al., 2016, 2017). However, because the intervention is not allocated at random, unknown differences between experimental and control groups may bias quasi-experiments' results (e.g. Michalopoulos, Bloom & Hill 2004). This problem, known as unobserved heterogeneity, historically led many in development economics to question their usefulness (e.g. Leamer 1983; also Levitt & List 2009; Angrist & Pischke 2010).

Randomised Control Trials ('RCTs'; also referred to as Randomised Controlled Trials) offer an outwardly straightforward solution to the limitations of other approaches to impact evaluation. By randomly allocating from the population of interest those units (individuals, areas or communities) which will receive a particular intervention (the 'treatment group'), and those which will not (the 'control group'), there should be no substantial differences in the types of unit that are in the treatment group when compared with the control group (e.g. White 2013). Evaluators can therefore assume that in the absence of the intervention, the outcomes of interest would have changed in the same way in the two groups making the control group a valid counterfactual for measuring the effect of the intervention can be calculated. Complete balance in all characteristics between treatment and control groups can only be guaranteed with extremely large sample sizes (e.g. Bloom 2008). However baseline data collection, stratification, and checking for balance between treatment and control groups can greatly reduce the probability of unbalanced groups (Glennerster & Takavarasha, 2013) and if differences remain this can be resolved through its inclusion as a covariate in subsequent analyses (Senn 2013). In any program, there may be a difference between the units which were potentially exposed to the intervention (all units in the treatment group) and those actually exposed (a sub-set of the intervention group). This arises because many interventions are voluntary and take-up will not be 100%, or units may fail to comply or drop out for many reasons. Evaluators therefore often calculate both the mean effect on units in the intervention group as a whole (the 'intention to treat') and the effect of the actual intervention on a treated unit (the 'treatment on the treated', e.g. Glennerster & Takavarasha 2013).

The relative simplicity and intuitiveness of RCTs may make them particularly appealing to policymakers, especially when compared with the statistical 'black box' of quasi-experiments, and this may make them more persuasive than other impact evaluation methods to sceptical audiences (Banerjee, Chassang & Snowberg, 2016). While the different kinds of quasi-experiment have associated with each of them a large number of assumptions in order for the counterfactual to be valid, and indeed the validity of the effect size estimate for any such quasi-experiment may be dependent upon the extent to which those assumptions are met, experimental evaluations such as RCTs avoid many of these problems and thus in some ways are conceptually simpler than quasi-experiments (Glennerster & Takavarasha, 2013). RCTs are also substantially less dependent on any theoretical understanding of *how* the intervention might or might not work.

RCTs are central to the paradigm of evidence-based medicine: since the 1940s tens of thousands of RCTs have been conducted and they are often considered the 'gold standard' for testing treatments' efficacy (Barton, 2000). They are also widely used in agriculture, education, social policy (Bloom, 2008), labour economics (List & Rasul, 2011), and, increasingly over the last two decades, in development economics (Banerjee & Duflo, 2011; Glennerster & Takavarasha, 2013). The governments of both the United Kingdom and the United States have strongly supported the use of RCTs in evaluating policy effectiveness (Haynes et al., 2012; Council of Economic Advisers, 2014). The United States Agency for International Development explicitly states that experimental impact evaluation provides the strongest evidence, and alternative methods should be used only when random assignment is not feasible (USAID, 2016). However there are both philosophical (e.g. Cartwright 2010) and practical (Deaton, 2010; Deaton & Cartwright, 2016) critiques of RCTs' use, and their recent spread in development economics has led to a polarized debate (e.g. Ravallion 2009; Picciotto 2012). This debate notwithstanding, some development RCTs have acted as a catalyst for the widespread implementation of interventions. A now classic RCT testing treatment of parasitic worm infection on health and educational outcomes in Kenyan schoolchildren (Miguel & Kremer, 2004) has led to the creation of initiatives such as Deworm the World (http://www.evidenceaction.org/dewormtheworld/) and the consequent treatment of over 95 million children.

Calls for the use of RCTs in evaluating environmental interventions have been increasing (Greenstone & Gayer, 2009; Pattanayak, 2009; Miteva, Pattanayak & Ferraro, 2012; Samii et al., 2014; Ferraro & Hanauer, 2014; Baylis et al., 2016; Curzon & Kontoleon, 2016; Börner et al., 2016, 2017). Many kinds of conservation interventions aim to deliver ecological outcomes through changing human behaviour through incentive structures or rules (e.g. agri-environment schemes, provision of alternative livelihoods, protected areas, payments for ecosystem services, and certification schemes). We term these *socio-ecological interventions*. There are clear lessons to be learnt from RCTs in development economics, which also aim to achieve development outcomes through changing human behaviour and therefore face similar issues. A few pioneering RCTs of such large-scale socio-ecological interventions have recently been concluded, evaluating: an incentive-based conservation program in Bolivia (described in this article; also see Grillos [2017] and Bottazzi et al. [2018]); a payment program for forest carbon in Uganda (Jayachandran et al., 2017); and unconditional cash transfers in support of conservation in Sierra Leone (Kontoleon et al., 2016). We expect that RCT evaluation in conservation will become more widespread in the coming years.

We examine the potential of RCTs in developing the evidence base supporting (or otherwise) use of conservation interventions and thereby supporting evidence-informed decision making. We discuss the factors influencing the usefulness, feasibility, and quality of RCT evaluation of conservation and aim to provide insights for researchers and practitioners interested in conducting high-quality evaluations. The structure of the chapter is mirrored by a checklist (figure 1) which can be used to assess the feasibility of an RCT in a given context. We also illustrate these points throughout the chapter with the implementation of the recent RCT of the incentive-based conservation program *Watershared* by the NGO *Fundación Natura Bolivia* (*Natura*) in Bolivia (figures 2 and 3).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26929v1 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

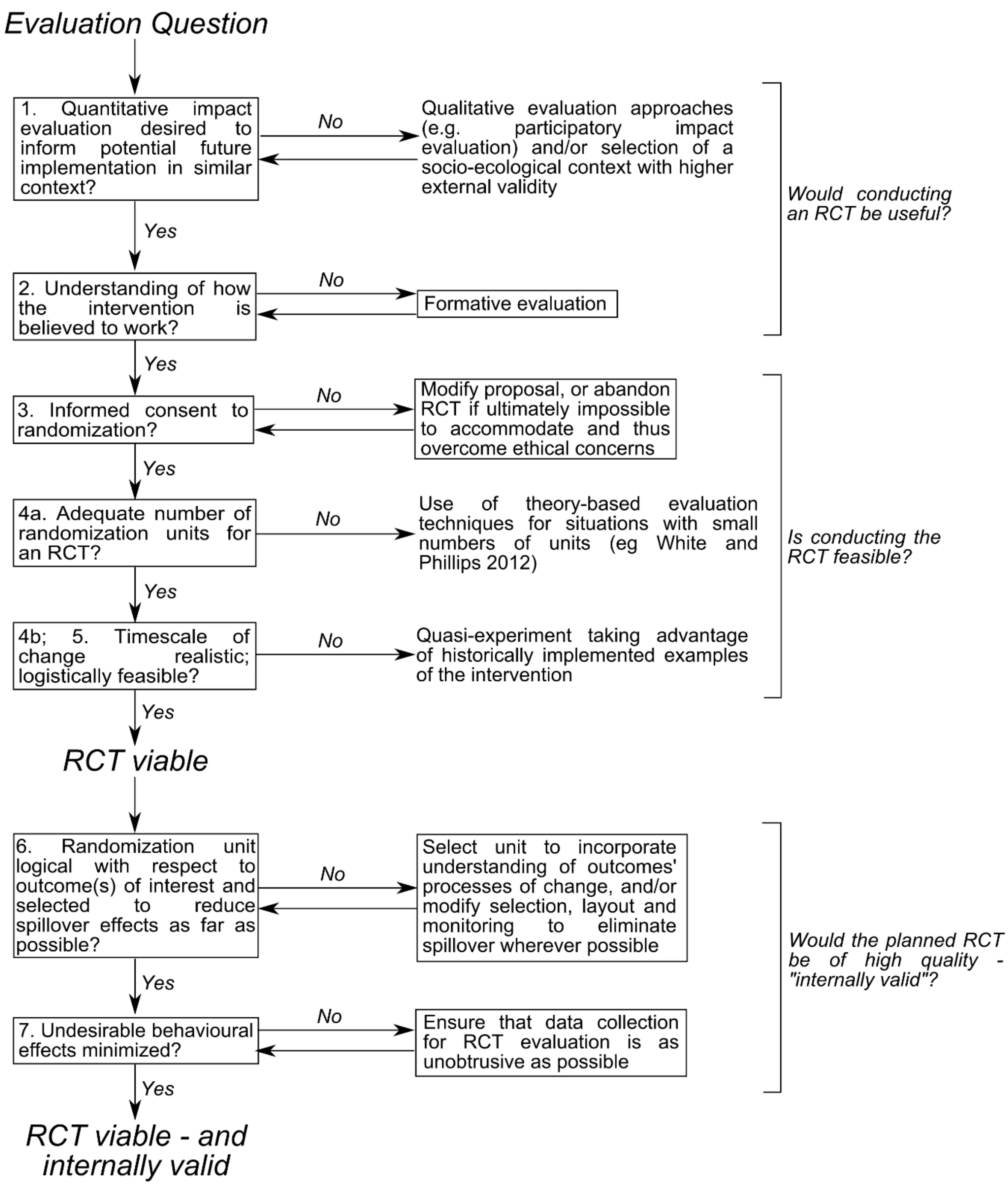

Figure 1. Summary of our suggested decision-making process for evaluators relating to RCT feasibility and quality, and alternative evaluation options if RCTs are inappropriate. Decisions or actions for evaluators to take during the process of RCT design are in boxes. Pattanayak (2009), Stern *et al.* (2012) and White & Phillips (2012) are good introductions to the alternative evaluation methods mentioned.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26929v1 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

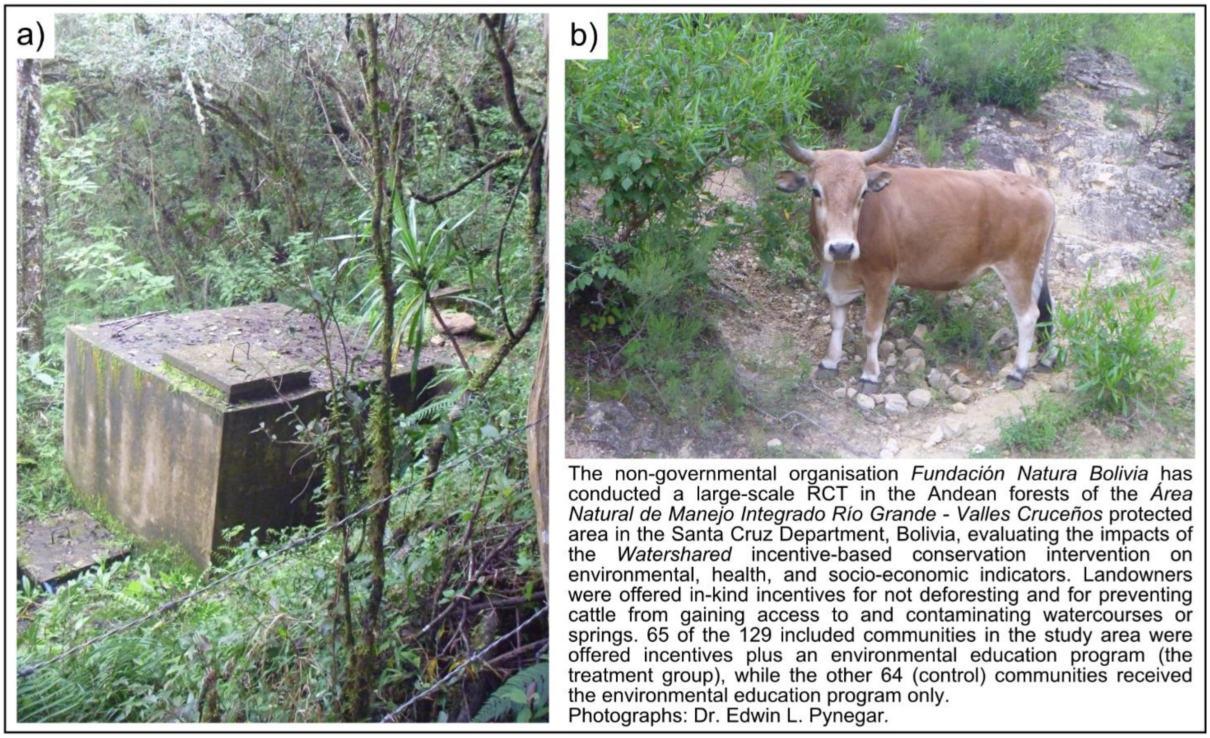

Figure 2. The Bolivian NGO *Fundación Natura Bolivia* conducted an RCT of their PES-like conservation program, *Watershared*, in the Bolivian Andes between 2011 and 2016. a) Water source located in forested land fenced off to prevent livestock access. b) Free-roaming cattle are common in the area and are widely seen as responsible for contaminating water supplies and degrading forests.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26929v1 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

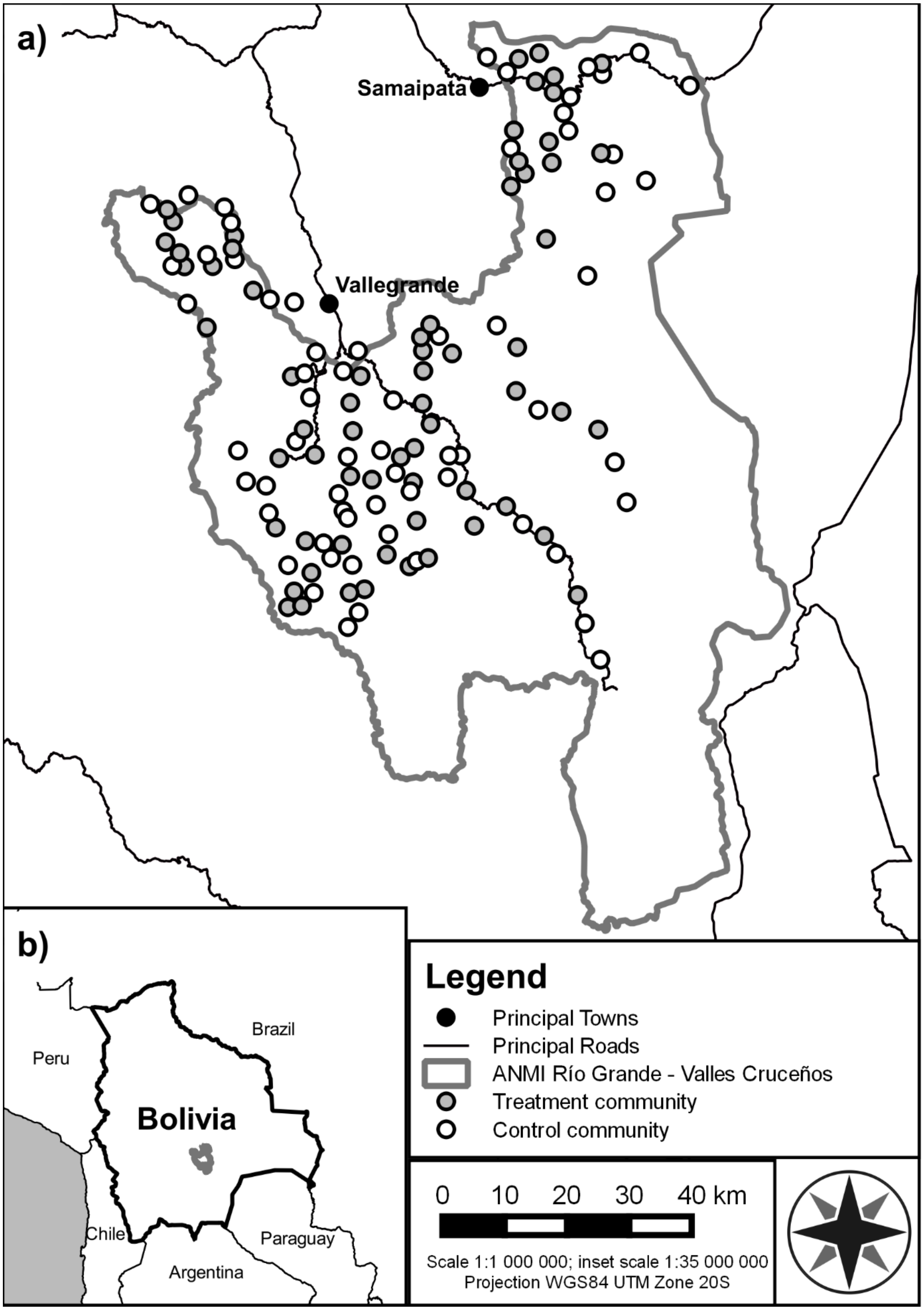

Figure 3. a) Locations of the 65 treatment and 64 control communities included in the RCT. b) Location of the *ANMI Río Grande – Valles Cruceños* protected area within Bolivia.

Under what circumstances might an RCT evaluation be useful?

RCTs quantitatively evaluate an intervention's impact in a particular context

Many different approaches can be used to evaluate an intervention's impact. We focus on quantitative approaches, which allow the magnitude of the effect of an intervention on outcomes of interest to be estimated, as is often required by policy makers. However, evaluators should bear in mind that more qualitative approaches such as participatory or theory-based impact evaluation methods (e.g. Stern *et al.* 2012) might be more suitable in cases where the intervention was implemented in very few units (White & Phillips, 2012) or when evaluators seek a detailed understanding of the pathways of change from intervention through to outcome (Cartwright, 2010). RCT results indicate *whether* an intervention works and to what extent, but policymakers may also wish to know *why* it works, to allow prediction of project success in other contexts.

This issue of *external validity* – the extent to which knowledge obtained can be generalized to other contexts – is a major focus of the debate surrounding RCT use in development economics (e.g. Deaton 2010; Cartwright 2010). Advocates for RCTs accept such critiques as partially valid (White, 2013), but note that RCTs provide complementary, not contradictory knowledge to other approaches to impact evaluation. Additionally the question of whether learning obtained in one location or context can be applicable to another is an epistemological question common to much applied research and is not limited to RCTs (Glennerster & Takavarasha, 2013).

Solutions to the external validity problem include conducting qualitative studies alongside an RCT (researchers will inevitably develop an understanding of the causal processes involved anyway), or using covariates to explore which factors influence outcome. The most obvious solution, however, is to conduct RCTs of the same kind of intervention in different socio-ecological contexts (White, 2013). While this is challenging due to the spatial and temporal scale of RCTs evaluating socio-ecological interventions, a number of groups of researchers have recently undertaken RCTs of incentive-based conservation programs (Kontoleon *et al.* 2016; Jayachandran *et al.* 2017; as well as the RCT described in this thesis). A study consisting of six separate RCTs on three continents, with over 10,000 participants in total, which evaluated a multifaceted development approach targeted at extremely poor households (Banerjee et al., 2015), has shown that multiple simultaneous RCTs of an intervention can be conducted (and in this case the pattern of lasting positive effects on income and assets was found across all countries).

In Bolivia, the NGO *Natura* wished to evaluate quantitatively the effects of the *Watershared* intervention (an incentive-based Payment for Ecosystem Services-like program) on water quality, biodiversity indicator species, deforestation rates, and human wellbeing. Similar socio-ecological systems exist throughout Latin America and incentive-based forest conservation projects have been widely implemented in montane forested regions. *Natura* is currently undertaking a complementary RCT of the intervention in the drier Bolivian Chaco (where land is held communally by indigenous people) and is in the process of designing a third, in a different part of the Chaco, which will evaluate, amongst other questions, the relative effectiveness of framing the intervention as a Payments for Ecosystem Services program or as a reciprocal agreement on its eventual outcomes. Additionally, in follow-up surveys at the end of the evaluation period, researchers have also extensively used qualitative methods to understand more profoundly processes of change within treatment communities (Bottazzi et al., 2018).

RCTs are likely most usefully conducted when the intervention is well developed

Impact evaluation is a form of summative evaluation (Scriven, 1967), meaning that it involves measuring outcomes. This can be contrasted with formative evaluation, which develops and improves the design of an intervention. Many evaluation theorists recommend a cycle of formative and summative evaluation, by which interventions may progressively be understood, refined, and evaluated (Rossi, Lipsey & Freeman, 2004). This is similar to the thinking behind adaptive management (Lindenmayer & Likens, 2009). Summative evaluation alone is somewhat inflexible as once started, aspects of the intervention cannot be changed. The substantial investment of time and resources in an RCT is therefore likely to be most appropriate when implementers are confident that they have an intervention whose functioning is reasonably well developed and understood (Pattanayak, 2009; Cartwright, 2010). Again, outputs from formative and summative evaluation represent complementary and not contradictory knowledge.

In Bolivia, *Natura* has been undertaking incentive-based forest conservation in the Bolivian Andes since 2003, and cattle exclusion from water sources had been conducted in the region for decades by another NGO and by local communities. Lessons learnt from these experiences were integrated into the design of the *Watershared* intervention as evaluated by the RCT which began in 2010.

What affects the feasibility of RCT evaluation?

Ethical challenges

Randomisation involves withholding the intervention from the control group so the decision to randomize is not a morally neutral one. A central ethical principle in medical RCTs is that to justify a randomised experiment, there must be significant uncertainty surrounding whether the treatment is in fact better than the control (a principle known as equipoise). The mechanisms through which an environmental intervention is intended to result in changes are often complex and poorly understood, meaning that in environmental RCTs there may indeed be uncertainty about whether the treatment is better than the control. Additionally, it is unclear whether obtaining equipoise should even always be an obligation for evaluators (e.g. Brody 2012), as how well – not just whether – an intervention works, and how cost-effective it is, are also important results for policymakers. It may be argued that lack of availability of high-quality evidence leading to resources being wasted on ineffective or only modestly effective interventions is also unethical (List & Rasul, 2011). Decisions such as these are not solely for researchers to make and must be handled with sensitivity (White, 2013).

Another central principle of research ethics states that no-one should be a participant in research without giving their free, prior and informed consent. Depending on the scale at which the intervention under evaluation is implemented, it may not be possible to obtain consent from every individual in an area. This can be overcome by randomising by community or administrative unit (not by individual) and then giving individuals the opportunity of opting into or out of the offered intervention. This may result in challenges for interpretation as the level at which the intervention is implemented (the individual) is different from the level at which the randomisation is conducted.

In Bolivia, the complex nature of the socio-ecological system, and the lack of initial understanding of the ways in which the intervention might affect or not affect it, meant there was real uncertainty about the effectiveness of *Watershared* on outcomes of interest. However, had monitoring shown immediate significant improvements in water quality in the experimental communities, *Natura* would have stopped the RCT and immediately implemented the intervention in all communities. Consent was granted by community leaders for the randomisation and individual households could choose to join the program or not.

Spatial and temporal scale

Larger numbers of randomisation units in an RCT allow reliable detection of smaller effect sizes (Bloom, 2008). This is easily achievable in small-scale experiments, such as those studying the effects of nest boxes on bird abundance or of wildflower verges on farmland invertebrate biodiversity; such trials have been a mainstay of applied ecology for decades (c.f. Fisher 1935). However, increases in scale of the intervention will make RCT implementation more challenging. A large randomisation unit (such as a protected area) will mean few available randomisation units, increasing the effect size required for a result to be statistically significant and decreasing the experiment's power (Bloom, 2008; Glennerster & Takavarasha, 2013). Large randomisation units are also likely to increase costs and logistical difficulties. However we emphasise that this does not make such evaluations impossible; two recent RCTs of a purely ecological intervention – impact of use of neonicotinoid-free seed on bee populations – were conducted across a number of sites throughout northern and central Europe (Rundlöf et al., 2015; Woodcock et al., 2017). When the number of units available is extremely small, RCTs will clearly not be possible and evaluation methods based upon expected theories of change may be more appropriate (White & Phillips, 2012).

For some interventions, measurable changes in outcomes may take years or even decades, due to long life cycles of relevant species and the slow and stochastic nature of many ecosystem changes. It is unlikely to be realistic for researchers or practitioners to set up and monitor RCTs over such timescales. In these cases RCTs are likely to be an inappropriate means of impact evaluation, and the best option for evaluators would likely consist of a well-designed quasi-experiment taking advantage of a historically implemented example of the intervention.

In the Bolivian case, an RCT of the *Watershared* intervention was feasible as the intervention units are relatively small (communities of 2 to 185 households) and baseline data allowed stratified random allocation of 129 communities to control or treatment. The RCT was run over 5 years (2011-2016). Effects on water quality should be observable over this timescale as cattle exclusion may result in decreases in waterborne bacterial concentration in under 1 year (Meals, Dressing & Davenport, 2010). However impacts on biodiversity may be expected to take substantially longer.

Available resources

RCTs require substantial human, financial and organizational resources for their design, implementation, monitoring, and subsequent evaluation. These resources are over and above the additional cost of monitoring in control units, because RCT design, planning, and the subsequent analysis and interpretation require substantial effort. USAID advises that a minimum of 3% of a project or program's budget be allocated to external evaluation (USAID, 2016), while the World Health Organization recommends 3-5% (WHO, 2013). The UN's Evaluation Group has noted that the sums allocated within the UN in the past cannot achieve robust impact evaluations without major uncounted external contributions (UNEG Impact Evaluation Task Force, 2013). Conducting a high-quality RCT is certainly not cheap; many conservation practitioners are already well aware of this (Curzon & Kontoleon, 2016).

Collaborations between researchers (with independent research funding) and practitioners (with a part of their program budget allocated to evaluation) can be an effective way for high quality impact evaluation to be conducted. This was the case with the evaluation of *Watershared* in Bolivia*:* the NGO had funding for implementation of the intervention from development and conservation organizations while the additional costs of the RCT came from research grants and collaborations with universities. Additionally, there are a number of organizations whose goals include conducting and funding high-quality impact evaluations (including RCTs), such as Innovations for Poverty Action (www.poverty-action.org), the Abdul Latif Jameel Poverty Action Lab (J-PAL; www.povertyactionlab.org), and the International Initiative for Impact Evaluation (3ie; www.3ieimpact.org).

What factors affect the quality – the 'internal validity' – of an RCT evaluation?

Potential for 'spillover', and how selection of randomisation unit may affect this

Evaluators must decide upon the unit at which allocation of the intervention is to occur. In medicine the unit is normally the individual, although some interventions may be allocated to groups. In development economics units may be individuals, households, schools, communities, or other groups while in conservation units could also potentially include fields, farms, habitat patches, protected areas, or others. Units selected should, however, logically correspond to the process of change by which the intervention is understood to lead to the desired outcome (Glennerster & Takavarasha, 2013).

In conservation RCTs, surrounding context will often be critical to interventions' functioning. This is also true of some RCTs in medicine or development economics, and hence evaluators can learn from these fields. Spatial context means that evaluators need to consider the potential for outcomes to 'spill over' between units – with positive effects from the intervention in treatment units affecting control units, or vice versa (Glennerster & Takavarasha, 2013; Baylis et al., 2016). It is easy to imagine species of interest moving from one unit to another because of habitat connectivity or water flowing down from a treatment area to a control one. These kinds of spillover, which we refer to as *biophysical* as they relate to ecological processes, thus cause changes achieved in treatment areas to affect outcomes of interest in control areas and thus reduce an intervention's apparent effect size. If an intervention were to be implemented in all areas rather than solely treatment areas (presumably the ultimate goal for practitioners), such effects would not occur. Spillover is particularly likely to occur if the randomisation unit and the natural unit of the intended ecological process of change do not align, meaning in practice the intervention would be implemented in areas which would affect outcomes at control sites, and vice versa.

Spillover effects are thus a property of the trial itself, and are recognized as important in some situations in development economics. For example, the influential RCT investigating treatment of worm infection in Kenyan schoolchildren used schools as the randomisation unit as children in the same school are likely to interact and re-infect each other more frequently than with children at other schools. It was explicitly designed to allow measurement of spillover (Miguel & Kremer, 2004); and showed (notwithstanding the re-analysis by Davey *et al.* [2015]) that deworming in treatment schools resulted in decreased worm burden in children attending nearby non-treatment schools. Such spillover also affected one of the very few attempts to conduct a large-scale environmental management RCT: the UK Government's RCT of badger culling in south-western England (Donnelly et al., 2005).

Preliminary consideration of spatial relationships between units, and the relationship between randomisation units and the process of change for the indicators, is critical for reducing or eliminating spillover and thus successfully undertaking internally valid conservation RCTs. Spillover may also be reduced by selecting indicators and/or sites to monitor which would still be relevant and meaningful but would be unlikely to suffer from spillover (such as by choosing a species to monitor with a small range size, or ensuring that a control area's monitoring site would not be directly downstream of a treatment area's in an RCT of a payments for watershed services program).

In the evaluation of *Watershared,* it proved difficult to select a randomisation unit that was politically feasible and worked for all outcomes of interest. *Natura* used the community as the randomisation unit as it would have been extremely difficult to have offered *Watershared* agreements to some members of communities and not to others. Community boundaries thus had to be drawn (these did not previously exist) and these did not always align well with area of land in the catchment of the communities' water sources. Thus while *Natura* did all it could to ensure that no community water quality monitoring site was directly downstream of another, land under conservation agreements in one community would sometimes be located in the catchment upstream of the monitoring site of another, risking biophysical spillover. The extent to which this spillover took place, and its consequences, can be studied empirically.

Consequences of human behavioural effects on evaluation of socio-ecological interventions

There is a key difference between *ecological* interventions that aim to have a direct impact on an ecosystem and *socio-ecological* interventions which seek to deliver ecosystem changes by changing human behaviour. Medical RCTs are generally double-blinded so neither the researcher nor the participants know who has been assigned to the treatment or control group. Double-blinding is possible for some ecological interventions such as pesticide impacts on non-target invertebrate diversity in an agroecosystem: implementers do not have to know whether they are applying the pesticide or a control. This was partially achieved in the large-scale study of neonicotinoids cited above (Rundlöf et al., 2015). However, it is harder to carry out double-blind trials of the effects of socio-ecological interventions, as the intervention's consequences can be observed by the researchers, and participants will know whether they are being offered the intervention or not.

Lack of blinding creates potential problems. Participants in control communities may observe activities in nearby treatment communities and implement aspects of them on their own, reducing the measured impact of the intervention. They may, however, also feel resentful at being excluded from a supposedly beneficial intervention and therefore reduce pre-existing pro-conservation behaviours (Alpízar et al., 2017). It may be possible to reduce or eliminate such phenomena through selecting units whose individuals infrequently interact with each other. Evaluators of the *Watershared* program in Bolivia were concerned that members of control communities might decide to protect watercourses themselves after seeing successful results elsewhere (which would be encouraging, suggesting local support for the intervention, but which would interfere with the evaluation by reducing the effect size of the intervention detected). They therefore included questions in their follow-up socio-economic surveys to identify this effect; these revealed only one case in over 1500 household surveys.

The second issue with lack of blinding is that RCT design is intended to achieve that treatment and control groups are not systematically different immediately after randomisation. However those allocated to control or treatment may have different expectations or show different behaviour or effort simply as a consequence of the awareness of being allocated to a control or treatment group, meaning that a systematic difference between the two groups would have been introduced (Chassang, Padró i Miquel & Snowberg, 2012). Hence the outcome observed may not depend solely on the efficacy of the intervention; some authors have claimed that these effects may be large (Bulte et al., 2014).

Overlapping terms have been introduced into the literature to describe the ways in which actions of participants in experiments vary due to differences in effort between treatment and control groups (summarised in table 1). The 'Hawthorne effect' describes the phenomenon that participants in an experiment may behave differently because they know that they are being studied (e.g. Levitt & List 2011). The 'Pygmalion' and 'golem' effects, in which participants may adjust effort to meet experimenter expectations, are a form of this (Babad, Inbar & Rosenthal, 1982). Similarly, treatment-group interviewees may give answers that they believe evaluators wish to hear, known as experimenter demand. The related 'John Henry effect' may arise when individuals in control groups increase effort to compete with the treatment group (Saretsky, 1972). In addition, it is rational for subjects to increase effort expended on implementing an intervention if they believe the intervention to be effective (Chassang, Padró i Miquel & Snowberg, 2012). The consequence of these 'rational effort' effects can be that performance increases when people believe in the intervention (Babad, Inbar & Rosenthal, 1982). Therefore, if an intervention appears to achieve a large change in an outcome of interest, that may be because true efficacy of the intervention was large, or because participants *believed* it to be large and thus expended large amounts of effort on implementing it.

We do not believe that potential behavioural effects invalidate RCT evaluation as some have claimed (Scriven, 2008), as part of an intervention's impact in subsequent implementation will also be due to implementers' expended effort (Chassang, Padró i Miquel & Snowberg, 2012). It remains unclear whether behavioural effects are large enough to result in incorrect inference, or even exist at all (Bausell, 2015). In the case of the evaluation of *Watershared*, compliance monitoring is an integral part of incentive-based or conditional conservation, so any behavioural effect driven by increased monitoring should be thought of as an effect of the intervention itself rather than a confounding influence on outcome. Any such effects may be reduced through low-impact monitoring (Glennerster & Takavarasha, 2013). In Bolivia, water quality measurement was unobtrusive (few community members were aware of *Natura* technicians being present) and infrequent (either annual or biennial); deforestation monitoring was even less obtrusive as it was based upon satellite imagery; and socio-economic surveys were undertaken equally in treatment and control communities.

Conclusions

Scientific evidence supporting an intervention's use does not necessarily lead to the uptake of that intervention. Policy is at best *evidence-informed* rather than *evidence-based* (Adams & Sandbrook, 2013) because cost and political acceptability inevitably influence decisions, and frameworks to integrate evidence into decision-making are often lacking (Segan et al., 2011). However, improving available knowledge of intervention effectiveness is still important. For example, managers are more likely to report an intention to change their management strategies when presented with high-quality evidence of intervention effectiveness (Walsh, Dicks & Sutherland, 2015). The potential for evidence to have influence is higher when it is driven by the needs of practitioners: links between researchers and policymakers or practitioners throughout the design and implementation of impact evaluation studies are therefore valuable (Cook et al., 2013).

RCTs can be used to establish a reliable counterfactual allowing robust estimation of intervention effectiveness, and hence cost-effectiveness, and interest in their use is increasing within the conservation community. Like any evaluation method, they are clearly not suitable in all circumstances, and there exist significant practical challenges with their implementation. Even when feasible, evaluators must design RCTs with great care to avoid spillover and behavioural effects and thus maintain internal validity. We would argue that it still remains unclear whether, to what extent, and in which contexts, RCTs are likely to provide estimates of treatment effects more accurate than quasi-experiments (c.f. Michalopoulos, Bloom & Hill 2004; Bulte *et al.* 2014), due to confounding experimental effects. This research question deserves a great deal more attention. There also will inevitably remain some level of subjectivity whether a location or context for subsequent implementation of an intervention is similar enough to one where an RCT was carried out to allow the learning to be confidently applied. We hope that those interested in evaluating the impact of conservation interventions can avoid the polarization and controversy surrounding their use in development economics while learning from their implementation in other fields. RCTs may then make a substantial contribution towards building a more robust evidence base to underpin conservation decisions.

Table 1. Consequences of behavioural effects when compared with results obtained in a hypothetical double-blind RCT. Hawthorne '1', '2' and '3' refer to the three kinds of effect discussed in Levitt & List (2011). References: [a] - (Jakovljevic, 2014). [b] - (Rosenthal & Jacobson, 1968). [c] - (Babad, Inbar & Rosenthal, 1982). [d] - (Levitt & List, 2011). [e] - (Orne, 1962).

Effect name	Description/Explanation	Other names	Effect on outcome in treatment units	Effect on outcome in control units	Effect on estimated effect size of intervention
'Hawthorne 1'	Act of observation increases effort	-	Increases	Increases	Unknown
'Hawthorne 2'	Changes in intervention increase effort	Halo effect of uncontrolled novelty[a]	None / Increases	None	None / Increases
'Hawthorne 3'	Experimental subjects tend to meet what they believe to be experimenters' expectations	Pygmalion effect[b]; golem effect[c]; Rosenthal effect[a]; experimenter demand[d]; demand characteristics[e]	Increases	None / Decreases	Increases
Rational effort	Experimental subjects base effort on their own expectations of the intervention's effectiveness	Galatea effect[c]	Increases	None / Decreases	Increases
'John Henry'	Individuals in control group increase effort in an attempt to compete with the intervention group	-	None	None / Increases	None / Decreases

Reference List

Adams WM., Sandbrook C. 2013. Conservation, evidence and policy. *Oryx* 47:329–335. DOI: 10.1017/S0030605312001470.

Alpízar F., Nordén A., Pfaff A., Robalino J. 2017. Spillovers from targeting of incentives: Exploring responses to being excluded. *Journal of Economic Psychology* 59:87–98. DOI: 10.1016/j.joep.2017.02.007.

Andam KS., Ferraro PJ., Pfaff A., Sanchez-Azofeifa GA., Robalino JA. 2008. Measuring the effectiveness of protected area networks in reducing deforestation. *Proceedings of the National Academy of Sciences of the United States of America* 105:16089–16094. DOI: 10.1073/pnas.0800437105.

Angrist JD., Pischke J-S. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives* 24:3–30. DOI: 10.1257/jep.24.2.3.

Babad EY., Inbar J., Rosenthal R. 1982. Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology* 74:459–474. DOI: 10.1037/0022-0663.74.4.459.

Banerjee A., Chassang S., Snowberg E. 2016. *Decision Theoretic Approaches to Experiment Design and External Validity*. NBER Working Paper No. 22167, Cambridge, MA. DOI: 10.3386/w22167.

Banerjee A., Duflo E. 2011. *Poor Economics*. New York: PublicAffairs.

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* 348:1260799. DOI: 10.1126/science.1260799.

Barton S. 2000. Which clinical studies provide the best evidence? *BMJ* 321:255–256. DOI: 10.1136/bmj.321.7256.255.

Bausell RB. 2015. *The Design and Conduct of Meaningful Experiments Involving Human Participants: 25 Scientific Principles*. New York: Oxford University Press.

Baylis K., Honey-Rosés J., Börner J., Corbera E., Ezzine-de-Blas D., Ferraro PJ., Lapeyre R., Persson UM., Pfaff A., Wunder S. 2016. Mainstreaming Impact Evaluation in Nature Conservation. *Conservation Letters* 9:58–64. DOI: 10.1111/conl.12180.

Bloom HS. 2008. The Core Analytics of Randomized Experiments for Social Research. In: Alasuutari P, Bickman L, Brannen J eds. *The SAGE Handbook of Social Research Methods*. London: SAGE Publications Ltd, pp. 115–133. DOI: http://dx.doi.org/10.4135/9781848608429.n9.

Börner J., Baylis K., Corbera E., Ezzine-de-Blas D., Ferraro PJ., Honey-Rosés J., Lapeyre R., Persson UM., Wunder S. 2016. Emerging Evidence on the Effectiveness of Tropical Forest Conservation. *PLOS ONE* 11:e0159152. DOI: 10.1371/journal.pone.0159152.

Börner J., Baylis K., Corbera E., Ezzine-de-Blas D., Honey-Rosés J., Persson UM., Wunder S. 2017. The Effectiveness of Payments for Environmental Services. *World Development* 96:359–374. DOI: 10.1016/j.worlddev.2017.03.020.

Bottazzi P., Wiik E., Crespo D., Jones JPG. 2018. Payment for Environmental "Self-Service": Exploring the Links Between Farmers' Motivation and Additionality in a Conservation Incentive Programme in the Bolivian Andes. *Ecological Economics* 150:11–23. DOI: 10.1016/j.ecolecon.2018.03.032.

Brody H. 2012. A critique of clinical equipoise. In: Miller FG ed. *The Ethical Challenges of Human Research*. New York: Oxford University Press, pp. 199–216. DOI: 10.1093/acprof:osobl/9780199896202.003.0015.

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* 96:813–830. DOI: 10.1093/ajae/aau015.

Butsic V., Lewis DJ., Radeloff VC., Baumann M., Kuemmerle T. 2017. Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology* 19:1–10. DOI: 10.1016/j.baae.2017.01.005.

Cartwright N. 2010. What are randomised controlled trials good for? *Philosophical Studies* 147:59–70. DOI: 10.1007/s11098-009-9450-2.

Chassang S., Padró i Miquel G., Snowberg E. 2012. Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments. *American Economic Review* 102:1279–1309. DOI: 10.1257/aer.102.4.1279.

Cook CN., Mascia MB., Schwartz MW., Possingham HP., Fuller RA. 2013. Achieving conservation science that bridges the knowledge-action boundary. *Conservation Biology* 27:669–678. DOI: 10.1111/cobi.12050.

Council of Economic Advisers. 2014. Evaluation as a tool for improving federal programs. In: *Economic Report of the President, Together with the Annual Report of the Council of Economic Advisors*. Washington DC: U.S. Government Printing Office, pp. 269–298.

Curzon HF., 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. DOI: 10.1016/j.jenvman.2016.05.062.

Davey C., Aiken AM., Hayes RJ., Hargreaves JR. 2015. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology* 44:1581–1592. DOI: 10.1093/ije/dyv128.

Deaton A. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48:424–455. DOI: 10.1257/jel.48.2.424.

Deaton A., Cartwright N. 2016. *Understanding and Misunderstanding Randomized Controlled*

Trials. NBER Working Paper N. 22595, Cambridge, MA. DOI: 10.3386/w22595.

Donnelly CA., Woodroffe R., Cox DR., Bourne FJ., Cheeseman CL., Clifton-Hadley RS., Wei G., Gettinby G., Gilks P., Jenkins H., Johnston WT., Le Fevre AM., McInerney JP., Morrison WI. 2005. Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature* 439:843–846. DOI: 10.1038/nature04454.

Ferraro PJ., Hanauer MM. 2014. Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annual Review of Environment and Resources* 39:495–517. DOI: 10.1146/annurev-environ-101813-013230.

Ferraro PJ., Pattanayak SK. 2006. Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments. *PLoS Biology* 4:e105. DOI: 10.1371/journal.pbio.0040105.

Fisher RA. 1935. *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.

Glennerster R., Takavarasha K. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press. DOI: 10.2307/j.ctt4cgd52.

Greenstone M., Gayer T. 2009. Quasi-experimental and experimental approaches to environmental economics. *Journal of Environmental Economics and Management* 57:21–44. DOI: 10.1016/j.jeem.2008.02.004.

Grillos T. 2017. Economic vs non-material incentives for participation in an in-kind payments for ecosystem services program in Bolivia. *Ecological Economics* 131:178–190. DOI: 10.1016/j.ecolecon.2016.08.010.

Haynes L., Service O., Goldacre B., Torgerson D. 2012. *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*. London: UK Government Cabinet Office Behavioural Insights Team. DOI: 10.2139/ssrn.2131581.

Independent Evaluation Group. 2012. *World Bank Group Impact Evaluations: Relevance and Effectiveness*. Washington DC: World Bank Group.

Jakovljevic M. 2014. The placebo–nocebo response: Controversies and challenges from clinical and research perspective. *European Neuropsychopharmacology* 24:333–341. DOI: 10.1016/j.euroneuro.2013.11.014.

Jayachandran S., de Laat J., Lambin EF., Stanton CY., Audy R., Thomas NE. 2017. Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science* 357:267–273. DOI: 10.1126/science.aan0568.

Kontoleon A., Conteh B., Bulte E., List JA., Mokuwa E., Richards P., Turley T., Voors M. 2016. *The impact of conditional and unconditional transfers on livelihoods and conservation in Sierra Leone, 3ie Impact Evaluation Report 46.* New Delhi: International Initiative for Impact Evaluation.

Leamer EE. 1983. Let's take the con out of econometrics. *American Economic Review* 73:31–43. DOI: 10.2307/1803924.

Levitt SD., List JA. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review* 53:1–18. DOI: 10.1016/j.euroecorev.2008.12.001.

Levitt SD., List JA. 2011. Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. *American Economic Journal: Applied Economics* 3:224–238. DOI: 10.1257/app.3.1.224.

Lindenmayer DB., Likens GE. 2009. Adaptive monitoring: a new paradigm for long-term research and monitoring. *Trends in Ecology & Evolution* 24:482–486. DOI: 10.1016/j.tree.2009.03.005.

List JA., Rasul I. 2011. Field Experiments in Labor Economics. In: Ashenfelter O, Card D eds. *Handbook of Labor Economics*. Amsterdam: North Holland, pp. 104–228. DOI: 10.1016/S0169-7218(11)00408-4.

Margoluis R., Russell V., Gonzalez M., Rojas O., Magdaleno J., Madrid G., Kaimowitz D. 2001. *Maximum Yield? Sustainable Agriculture as a Tool for Conservation*. Washington DC: Biodiversity Support Program.

Margoluis R., Stem C., Salafsky N., Brown M. 2009. Design alternatives for evaluating the impact of conservation projects. *New Directions for Evaluation* 122:85–96. DOI: 10.1002/ev.298.

Meals DW., Dressing SA., Davenport TE. 2010. Lag time in water quality response to best management practices: a review. *Journal of Environmental Quality* 39:85–96. DOI: 10.2134/jeq2009.0108.

Michalopoulos C., Bloom HS., Hill CJ. 2004. Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? *Review of Economics and Statistics* 86:156–179. DOI: 10.1162/003465304323023732.

Miguel E., Kremer M. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72:159–217. DOI: 10.1111/j.1468-0262.2004.00481.x.

Miteva DA., Pattanayak SK., Ferraro PJ. 2012. Evaluation of biodiversity policy instruments: What works and what doesn't? *Oxford Review of Economic Policy* 28:69–92. DOI: 10.1093/oxrep/grs009.

Orne MT. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17:776–783. DOI: 10.1037/h0043424.

Pattanayak SK. 2009. *Rough Guide to Impact Evaluation of Environmental and Development Programs*. Kathmandu, Nepal: South Asian Network for Development and Environmental Economics.

Pattanayak SK., Wunder S., Ferraro PJ. 2010. Show me the money: Do payments supply environmental services in developing countries? *Review of Environmental Economics and Policy* 4:254–274. DOI: 10.1093/reep/req006.

Picciotto R. 2012. Experimentalism and development evaluation: Will the bubble burst? *Evaluation* 18:213–229. DOI: 10.1177/1356389012440915.

Pullin AS., Knight TM., Stone DA., Charman K. 2004. Do conservation managers use scientific evidence to support their decision-making? *Biological Conservation* 119:245–252. DOI: 10.1016/j.biocon.2003.11.007.

Rasolofoson RA., Ferraro PJ., Jenkins CN., Jones JPG. 2015. Effectiveness of Community Forest Management at reducing deforestation in Madagascar. *Biological Conservation* 184:271–277. DOI: 10.1016/j.biocon.2015.01.027.

Ravallion M. 2009. Should the Randomistas Rule? *The Economists' Voice* 6:8–12. DOI: 10.2202/1553-3832.1368.

Rosenthal R., Jacobson L. 1968. Pygmalion in the classroom. *The Urban Review* 3:16–20. DOI: 10.1007/BF02322211.

Rossi P., Lipsey M., Freeman H. 2004. *Evaluation: a Systematic Approach*. Thousand Oaks, CA: SAGE Publications.

Rundlöf M., Andersson GKS., Bommarco R., Fries I., Hederström V., Herbertsson L., Jonsson O., Klatt BK., Pedersen TR., Yourstone J., Smith HG. 2015. Seed coating with a neonicotinoid insecticide negatively affects wild bees. *Nature* 521:77–80. DOI: 10.1038/nature14420.

Samii C., Lisiecki M., Kulkarni P., Paler L., Chavis L. 2014. Effects of Payment for Environmental Services (PES) on Deforestation and Poverty in Low and Middle Income Countries: A Systematic Review. *Campbell Systematic Reviews* 10.

Saretsky G. 1972. The OEO PC experiment and the John Henry effect. *Phi Delta Kappan* 53:579–581.

Scriven M. 1967. The methodology of evaluation. In: Tyler RW, Gagne RM, Scriven M eds. *Perspectives of curriculum evaluation*. Chicago, IL: Rand McNally, pp. 39–83.

Scriven M. 2008. A summative evaluation of RCT methodology: and an alternative approach to causal research. *Journal of Multidisciplinary Evaluation* 5:11–24.

Segan DB., Bottrill MC., Baxter PWJ., Possingham HP. 2011. Using Conservation Evidence to Guide Management. *Conservation Biology* 25:200–202. DOI: 10.1111/j.1523-1739.2010.01582.x.

Senn S. 2013. Seven myths of randomisation in clinical trials. *Statistics in Medicine* 32:1439–1450. DOI: 10.1002/sim.5713.

Stern E., Stame N., Mayne J., Forss K., Davies R., Befani B. 2012. *Broadening the Range of Designs and Methods for Impact Evaluations*. London: UK Government Department for International Development.

Sutherland WJ., Pullin AS., Dolman PM., Knight TM. 2004. The need for evidence-based conservation. *Trends in Ecology and Evolution* 19:305–308. DOI: 10.1016/j.tree.2004.03.018.

UNEG Impact Evaluation Task Force. 2013. *Impact Evaluation in UN Agency Evaluation Systems: Guidance on Selection, Planning and Management*. New York: United Nations.

USAID. 2016. *Evaluation: Learning from Experience. USAID Evaluation Policy*. Washington DC: United States Agency for International Development.

Walsh JC., Dicks LV., Sutherland WJ. 2015. The effect of scientific evidence on conservation practitioners' management decisions. *Conservation Biology* 29:88–98. DOI: 10.1111/cobi.12370.

White H. 2013. An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness* 5:30–49. DOI: 10.1080/19439342.2013.764652.

White H., Phillips D. 2012. *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework*. New Delhi: International Initiative for Impact Evaluation.

WHO. 2013. *WHO Evaluation Practice Handbook*. Geneva, Switzerland: World Health Organization.

Woodcock BA., Bullock JM., Shore RF., Heard MS., Pereira MG., Redhead J., Ridding L., Dean H., Sleep D., Henrys P., Peyton J., Hulmes S., Hulmes L., Sárospataki M., Saure C., Edwards M., Genersch E., Knäbe S., Pywell RF. 2017. Country-specific effects of neonicotinoid pesticides on honey bees and wild bees. *Science* 356:1393–1395. DOI: 10.1126/science.aaa1190.