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

Pynegar, Edwin L.; Gibbons, James M.; Asquith, Nigel M.; Jones, Julia P.G.

**Oryx**

*DOI:*
10.1017/S0030605319000188

E-pub ahead of print: 24/10/2019

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

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

***Affiliations and Addresses:***

[a] College of Engineering and Environmental Sciences, 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: College of Engineering and Environmental Sciences, Bangor University, Deiniol Road, Bangor, Gwynedd LL57 2UW, UK

Email: edwin.pynegar@gmail.com

Edwin L. Pynegar ORCID: https://orcid.org/0000-0001-5975-696X

Julia P. G. Jones ORCID: https://orcid.org/0000-0002-5199-3335

## Abstract

The effectiveness of many widely used conservation interventions is poorly understood due to a lack of high-quality impact evaluations. Randomised Control Trials (RCTs), in which experimental units are randomly allocated to treatment or control groups, offer an intuitive means of calculating the impact of an intervention through establishing a reliable counterfactual scenario. As many conservation interventions depend on changing people's behaviour, conservation impact evaluation can learn a great deal from RCTs in fields such as development economics, where RCTs have become widely used but are controversial. We build on relevant literature from other fields to discuss how RCTs, despite their potential, are just one of a number of ways to evaluate impact, are not feasible in all circumstances, and factors such as spillover between units and behavioural effects must be considered in their design. We offer guidance and a set of criteria for deciding when RCTs may be an appropriate approach for evaluating conservation interventions, and factors to consider to ensure an RCT is of high quality. We illustrate this with examples from one of the very few concluded RCTs of a large-scale conservation intervention – that of an incentive-based conservation program in the Bolivian Andes. We argue that conservation should aim to avoid a re-run of the polarized debate surrounding the use of RCTs in other fields. RCTs will not be possible or appropriate in many circumstances, but if used carefully they can certainly be useful and could become a more widely used tool in the conservation impact evaluator's toolkit.

## Keywords

Counterfactual, Evidence, Effectiveness, Impact Evaluation, Randomization, Randomized Control Trials, RCTs.

## Introduction

It is widely recognised that conservation decisions should be evidence-informed (Pullin et al., 2004; Segan et al., 2011). Despite this, decisions often remain only weakly informed by the evidence base (e.g. Sutherland & Wordley, 2017). While this is at least partly due to decision makers' continuing lack of access to evidence (Rafidimanantsoa et al., 2018), complacency surrounding ineffective interventions (Pressey et al., 2017; Sutherland & Wordley, 2017), and perceived irrelevance of research to decision-making (Rose et al., 2018; Rafidimanantsoa et al., 2018), there are also limitations in the available evidence on the likely impacts of conservation interventions (Ferraro & Pattanayak, 2006; McIntosh et al., 2018). This has resulted in a growing interest in conservation impact evaluation (Ferraro & Hanauer, 2014; Baylis et al., 2016; Börner et al., 2016; Pressey et al., 2017), and to the creation of initiatives to facilitate access to and systematise the existing evidence, such as the Collaboration for Environmental Evidence (Anon., 2019a) and Conservation Evidence (Anon., 2019b).

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 (Miteva et al., 2012; Ferraro & Hanauer, 2014; Baylis et al., 2016; Pressey et al., 2017). It is well recognized that simple before-and-after comparison of units exposed to the intervention is flawed, as factors other than the intervention may have caused change in the outcomes of interest (Ferraro & Hanauer, 2014; Baylis et al., 2016). Simply comparing groups exposed and not exposed to the intervention is also flawed as the groups may differ in other ways that affect the outcome.

One solution is to replace post-project monitoring with more robust quasi-experiments, in which a variety of approaches may be used to construct a counterfactual scenario statistically (Glennerster & Takavarasha, 2013; Butsic et al., 2017). For example, matching involves comparing outcomes in units where an intervention is implemented with outcomes in similar units (identified statistically) which lack the intervention. This is increasingly used for conservation impact evaluations, such as determining the impact of national park establishment (Andam et al., 2008) or Community Forest Management (Rasolofoson et al., 2015) on deforestation. Quasi-experiments have a major role to play in conservation impact evaluation, and in some situations they will be the only robust option available to evaluators (Baylis et al., 2016; Butsic et al., 2017). However, because the intervention is not allocated at random, unknown differences between treatment and control groups may bias quasi-experiments' results (Michalopoulos et al., 2004; Glennerster & Takavarasha, 2013). This problem historically led many in development economics to question their usefulness (Angrist & Pischke, 2010). Each kind of quasi-experiment has associated assumptions which, if not met, affect the validity of the evaluation result (Glennerster & Takavarasha, 2013).

Randomised Control Trials ('RCTs'; also 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 which will receive a particular intervention (the 'treatment group'), and those which will not (the 'control group'), there should be no systematic differences between groups (White, 2013b). 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.

This relative simplicity of RCTs, especially when compared with the statistical black box of quasi-experiments, may make them more persuasive than other impact evaluation methods to sceptical audiences (Banerjee et al., 2016; Deaton & Cartwright, 2018). They are also – in theory – substantially less dependent than quasi-experiments on any theoretical understanding of how the intervention might or might not work (Glennerster & Takavarasha, 2013). RCTs are central to the paradigm of evidence-based medicine, and since the 1940s tens of thousands of RCTs have been conducted with them 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 (Ravallion, 2009; Banerjee et al., 2016; Leigh, 2018; Deaton & Cartwright, 2018). 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 exist both philosophical (e.g. Cartwright, 2010) and practical (Deaton, 2010; Deaton & Cartwright, 2018) critiques of RCTs. The statistical basis of randomised analyses is also not as simple as it might initially appear; randomisation can only be guaranteed to lead to complete balance between treatment and control groups with extremely large samples (Bloom, 2008). (However baseline data collection and stratification can greatly reduce the probability of unbalanced groups and remaining differences can be resolved through inclusion of covariates in analyses [Glennerster & Takavarasha, 2013]). Evaluators also often calculate both the mean effect on units in the treatment group as a whole (the 'intention to treat') and the effect of the actual intervention on a treated unit (the 'treatment on the treated'). These approaches will often give quite different results as there is commonly imperfect uptake of an intervention (a drug may not be taken correctly by all individuals in a treatment group, for example).

Regardless of the polarised debate that RCTs' spread in development economics has caused (Ravallion, 2009; Deaton & Cartwright, 2018), some development RCTs have acted as a catalyst for the widespread implementation of trialled interventions (Leigh, 2018). There are increasing calls for more use of RCTs in evaluating environmental interventions (Pattanayak, 2009; Miteva et al., 2012; Samii et al., 2014; Ferraro & Hanauer, 2014; Baylis et al., 2016; Curzon & Kontoleon, 2016; Börner et al., 2016, 2017). As many kinds of conservation program aim to deliver environmental improvements through changing human behaviour (e.g. agri-environment schemes, provision of alternative livelihoods, protected area establishment, payments for ecosystem services, REDD+ programs, and certification programs; we term these socio-ecological interventions), there are clear lessons to be learnt from RCTs in development economics, which aim to achieve development outcomes through changing behaviour.

A few pioneering RCTs of such socio-ecological interventions have recently been concluded (although these may not be fully exhaustive), evaluating: an incentive-based conservation program in Bolivia known as Watershared, described in this article; a payment program for forest carbon in Uganda (Jayachandran et al., 2017); unconditional cash transfers in support of conservation in Sierra Leone (Kontoleon et al., 2016); and a program aimed at reducing wild meat consumption in the Brazilian Amazon through social marketing and incentivising consumption of chicken (Chaves et al., 2018). We expect that RCT evaluation in conservation will become more widespread in the coming years.

We draw on a range of literature to examine the potential of RCTs for impact evaluation in the context of conservation. We discuss the factors influencing the usefulness, feasibility, and quality of RCT evaluation of conservation and aim to provide insights and guidance for researchers and practitioners interested in conducting high-quality evaluations. The structure of the article is mirrored by a checklist (Figure 1) which can be used to assess the suitability of an RCT in a given context. We illustrate these points with the recent RCT evaluating the Watershared incentive-based conservation program in the Bolivian Andes. This program, implemented by the NGO Fundación Natura Bolivia ('Natura'), aims to reduce deforestation, conserve biodiversity, and provide socio-economic and water quality benefits to local communities (Bottazzi et al., 2018; Pynegar et al., 2018; Wiik et al., 2019; Figure 2).

## Under what circumstances might an RCT evaluation be useful?

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

RCTs are a quantitative approach allowing the magnitude of the effect of an intervention on outcomes of interest to be estimated. Qualitative approaches based on causal chains or theory of change might be more suitable where such quantitative estimates are not needed or where the intervention can only be implemented in very few units (e.g. White & Phillips, 2012) or when the focus is on understanding the pathways of change from intervention through to outcome (Cartwright, 2010). Some have argued that such mechanistic understanding is more valuable than estimates of effect sizes for practitioners and policymakers (Cartwright, 2010; Miteva, Pattanayak & Ferraro, 2012; Deaton & Cartwright, 2018). To put this another way, RCTs can indicate whether an intervention works and to what extent, but policy makers often 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 from an RCT can be generalized to other contexts – is a major focus of the controversy surrounding RCT use in development economics (e.g. Deaton, 2010; Cartwright, 2010). Advocates for RCTs accept such critiques as partially valid (e.g. White, 2013b) and acknowledge that RCTs should be considered as providing complementary and not contradictory knowledge to other approaches. Firstly, more qualitative studies can be conducted alongside an RCT to examine processes of change; indeed most evaluators who advocate RCTs clearly also recognise that combining quantitative and qualitative approaches is likely to be most informative (e.g. White, 2013a). Secondly, researchers can use covariates to explore which contextual features affect outcomes of interest, to look for those features upon future implementation of the intervention (although to avoid data dredging, ideally hypotheses and analysis plans should be pre-registered). Statistical methods can also be used to explore heterogeneous responses within treatment groups in an RCT (Glennerster & Takavarasha, 2013), and RCTs may also be designed to answer more complex contextual questions through trials with multiple treatment groups or other modifications to the basic setup (Bonell et al., 2012). Thirdly, evaluators may conduct RCTs of the same kind of intervention in different socio-ecological contexts (White, 2013b), which increases results' generalisability. While this is challenging due to the spatial and temporal scale of RCTs evaluating socio-ecological interventions, researchers have recently undertaken a number of RCTs of incentive-based conservation programs (Kontoleon et al., 2016; Jayachandran et al., 2017; Pynegar et al., 2018). Finally, 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).

In the RCT evaluating the Bolivian Watershared program, the external validity issue has been addressed as a key concern. Similar socio-ecological systems exist throughout Latin America and incentive-based forest conservation projects have been widely implemented (Asquith, 2016). Natura is currently undertaking two complementary RCTs of the intervention in other parts of Bolivia. Finally, researchers used a combination of both qualitative and quantitative methods at the end of the evaluation period to understand in more depth participant motivation and processes of change within treatment communities (Bottazzi et al., 2018) as well as comparing outcomes in control and treatment communities (Pynegar et al., 2018; Wiik et al., 2019).

**RCTs are most usefully conducted when the intervention is reasonably well developed**

Impact evaluation is a form of summative evaluation, meaning that it involves measuring outcomes of an established intervention. This can be contrasted with formative evaluation, which progressively 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 et al., 2004), which is similar to the thinking behind adaptive management (McCarthy & Possingham, 2007; Gillson et al., 2019). Summative evaluation alone is inflexible as once started, aspects of the intervention cannot sensibly be changed (at least not without losing external validity). 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 understood (Pattanayak, 2009; Cartwright, 2010).

In Bolivia, Natura has been undertaking incentive-based forest conservation in the Bolivian Andes since 2003. Learning from these experiences was integrated into the design of the Watershared intervention as evaluated by the RCT which began in 2010. However, despite this substantial experience developing the intervention, there were challenges with its implementation in the context of the RCT which in retrospect affected both the program's effectiveness and the evaluation's usefulness. For example, uptake of the agreements was quite low (Wiik et al., 2019), and little of the most important land from a water quality perspective was enrolled in Watershared agreements. Given this low uptake, the lack of an observed effect of the program on water quality at landscape scale might have been predicted without the RCT (Pynegar et al., 2018). Further formative evaluation of uptake rates and likely spatial patterns of implementation before the RCT was implemented would have been valuable.

## What affects the feasibility of RCT evaluation?

**Ethical challenges**

Randomisation involves withholding the intervention from the control group, so the decision to randomise is not a morally neutral one. An ethical principle in medical RCTs is that to justify a randomised experiment, there must be significant uncertainty surrounding whether the treatment is better than the control (a principle known as equipoise; Brody, 2012). Experiments such as randomly allocating areas to be deforested or not to investigate ecological impacts would clearly not be ethical, which is why the Stability of Altered Forest Ecosystems project, for example, made use of already planned deforestation (Ewers et al., 2011). However the mechanisms through which many conservation interventions, especially socio-ecological interventions, are intended to result in change are often complex and poorly understood, meaning that in such RCTs there often will indeed be uncertainty about whether the treatment is better. Additionally, it is debatable whether obtaining equipoise should even always be an obligation for evaluators (e.g. Brody 2012), as how well an intervention works, and how cost-effective it is, are also

**Commented [JPGJ1]:** Just cite once in the sentence-its fine here important results for policymakers (White, 2013b). It may be argued that lack of availability of high-quality evidence leading to resources being wasted on ineffective interventions is also unethical (List & Rasul, 2011). Decisions such as these are not solely for researchers to make and must be sensitively handled (White, 2013b).

Another principle of research ethics states that no one should be a participant in an experiment without giving their free, prior and informed consent. Depending on the scale at which the intervention is implemented, it may not be possible to obtain consent from every individual in an area. This might be overcome by randomising by community rather than individual and then giving individuals in the treatment community the opportunity to opt into the intervention. This shows how implementers can think flexibly to overcome ethical challenges.

In Bolivia, the complex nature of the socio-ecological system, and the initial relative lack of understanding of the ways in which the intervention might affect it, meant there was genuine uncertainty about Watershared's effectiveness. However, had monitoring shown immediate significant improvements in water quality in treatment communities, *Natura* would have stopped the RCT and implemented the intervention in all communities. Consent was granted by mayors for the randomisation and individual landowners could choose to sign an agreement or not. While this was both more ethically acceptable and in reality the only way to implement Watershared agreements in this socio-ecological context, it led to variable (and sometimes low) uptake of the intervention, complicating the subsequent evaluation (Wiik et al., 2019).

**Spatial and temporal scale**

Larger numbers of randomisation units in an RCT allow detection of smaller significant 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 invertebrate biodiversity; such trials have been a mainstay of applied ecology for decades. However, increases in scale of the intervention will make RCT implementation more challenging. Interventions implemented at a large scale will likely have few randomisation units available for an RCT, 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 very low, however, RCTs will not be appropriate and theory-based evaluations based upon analysing expected theories of change may be more sensible (e.g. White & Phillips, 2012). Such theory-based evaluations allow attribution of changes in outcomes of interest to particular interventions, but do not allow estimation of treatment effect sizes.

For some conservation interventions, measurable changes in outcomes may take years or even decades, due to long life cycles of species or the slow and stochastic nature of many ecosystem changes. It is unlikely to be realistic 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 quasi-experiment taking advantage of a historically implemented example of the intervention.

In the Bolivian case, an RCT of the Watershared intervention was ambitious but feasible (129 communities as randomisation units, each consisting of 2 to 185 households). Following baseline data collection in 2010, the intervention was first offered in 2011 and endline data was collected in 2015-16. Effects on water quality were expected to be observable over this timescale as cattle exclusion can result in decreases in waterborne bacterial concentration in under 1 year (Meals et al., 2010). However Pynegar et al. (2018) did not find an impact of the intervention on water quality at landscape scale, and time-lags may be part of the reason for this. Neither did Wiik et al. (2019) find a strong impact of the program on deforestation. One hypothesis explaining this is that impacts may take longer to materialise as they can depend on the development of alternative livelihoods introduced as part of the program.

**Available resources**

RCTs require substantial human, financial and organisational resources for their design, implementation, monitoring and evaluation. These resources are above the additional cost of monitoring in control units, because RCT design, planning, and subsequent analysis and interpretation require substantial effort and knowledge. 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). As conservation practitioners are already aware, conducting a high-quality RCT is not cheap (Curzon & Kontoleon, 2016).

Collaborations between researchers (with independent funding) and practitioners (with a part of their program budget) can be an effective way for high-quality impact evaluation to be conducted. This was the case with the evaluation of Watershared*: Natura* had funding for implementation of the intervention from development and conservation organisations, while the additional costs of the RCT came from separate research grants. 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, the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology, and the International Initiative for Impact Evaluation (3ie).

## 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; 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 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. Outcomes may 'spill over' – with changes achieved by the intervention in treatment units affecting outcomes of interest in control units (Glennerster & Takavarasha, 2013; Baylis et al., 2016) – at least in cases where the randomisation unit is not 'closed' or somehow bounded in a way that prevents this from happening. For example, an RCT evaluating a successful community-based anti-poaching program would suffer from spillover if population increases in the treatment community-associated areas resulted in these acting as a source of individuals for control areas. Spillover thus reduces 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 spillover would not occur, and so it is a property of the trial itself. Such spillover affected one of the few large-scale environmental management RCTs: that evaluating badger culling in south-western England (Donnelly et al., 2005).

Spillover is particularly likely to occur if the randomisation unit and the natural unit of the intended ecological process of change are incongruent, meaning the intervention would inevitably be implemented in areas which would affect outcomes in control units. Therefore, consideration of spatial relationships between units, and of the relationship between randomisation units and the outcomes' process of change, is critical. For example the anti-poaching program described above might instead use closed groups or populations of the target species as the randomisation unit, with the program then implemented in communities covering the range of each treatment group. Spillover may also be reduced by selecting indicators and/or sites to monitor which would still be relevant but would be unlikely to suffer from it (i.e. more bounded units or monitoring sites – 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 RCT of Watershared, it proved difficult to select a randomisation unit that was politically feasible and worked for all outcomes of interest. Natura used community as the randomisation unit, so community boundaries had to be defined and these did not always align well with the watersheds supplying the communities' water sources. While very few water quality monitoring sites were directly downstream of another, land under agreements in one community would sometimes be located in the watershed upstream of the monitoring site of another, risking spillover. The extent to which this took place, and its consequences, were studied empirically(Pynegar, 2018) However, the randomisation unit worked well for the deforestation analysis. Communities have easily defined boundaries (although see Wiik et al., 2019) and offering the program by community was most practical logistically. A smaller unit would have presented issues of perceived fairness as it would have been extremely difficult to have offered Watershared agreements to some members of communities and not to others. Jayachandran et al. (2017)'s RCT also selected community as the randomisation unit.

**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 (see Rundlöf et al., 2015). However, it is harder to carry out double-blind trials of socio-ecological interventions, as the intervention's consequences can be observed by the evaluators (even if they are not the people actually implementing it) and participants will obviously know whether they are being offered the intervention.

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. Alternatively, they may feel resentful at being excluded from a beneficial intervention and therefore reduce 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 Watershared believed that members of control communities might decide to protect watercourses themselves after seeing successful results elsewhere (which would be encouraging for the NGO, suggesting local support for the intervention, but which would interfere with the evaluation by reducing the estimated intervention effect size). They therefore included questions in endline socio-economic surveys to identify this effect; these revealed only one case in over 1500 household surveys (Pynegar, 2018).

The second issue with lack of blinding is that randomisation 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 (Chassang et al., 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). We do not believe that behavioural effects inevitably invalidate RCT evaluation as some have claimed (Scriven, 2008), as part of any intervention's impact when implemented will be due to implementers' expended effort (Chassang et al., 2012). It also remains unclear whether behavioural effects are large enough to result in incorrect inference (Bulte et al., 2014; 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 rather than a confounding influence. Such effects may also be reduced through low-impact monitoring (Glennerster & Takavarasha, 2013). 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; Rose et al., 2018) 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, conservation managers are more likely to report an intention to change their management strategies when presented with high-quality evidence (Walsh et al., 2015). Conservation science therefore needs to use the best possible approaches for evaluation of interventions.

Like any evaluation method, Randomised Control Trials are clearly not suitable in all circumstances. Large-scale RCTs are unlikely to be a worthwhile approach to impact evaluation unless the intervention to be evaluated is well understood, either from theory or previous formative evaluation. Even when feasible and potentially useful, RCTs must be designed with great care to avoid spillover and behavioural effects. 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 learning to be confidently applied. However RCTs can be used to establish a reliable and intuitively plausible counterfactual and therefore provide a robust estimate of intervention effectiveness, and hence cost-effectiveness. It is therefore unsurprising that interest in their use is increasing within the conservation community. We hope that those interested in evaluating the impact of conservation interventions can learn from the use of RCTs in other fields while avoiding the polarisation and controversy surrounding them. Over time RCTs may then make a substantial contribution towards conservation impact evaluation.

## Author Contributions

Review of the relevant literature: ELP; writing of the article: ELP, JMG, NMA, JPGJ.

## Acknowledgements

This work was supported by a Doctoral Training Grant from the Natural Environment Research Council (1358260) and a grant from the Leverhulme Trust (RPG-2014-056). Nigel Asquith acknowledges a Charles Bullard Fellowship from the Harvard Forest, and grants NE/I00436X/1 and NE/L001470/1 from the Ecosystem Services for Poverty Alleviation (ESPA) Program. We are grateful to all our colleagues and collaborators at Fundación Natura Bolivia, particularly María Teresa Vargas and Tito Vidaurre, for valued discussion and for Jörn Scharlemann for helpful comments on the text. We would also like to thank two anonymous reviewers for their edits, suggestions, and help in improving and clarifying our thinking.

## Conflict of Interest Statement

ELP authored this review while an independently funded Ph. D. candidate, but since then has worked for Fundación Natura Bolivia in a consulting role. NMA worked for many years as the Director of Strategy and Policy at Natura and still has close personal relationships with staff at Natura.

## Ethical Standards

This research fully complies with the *Oryx* Code of Conduct. The research did not involve human subjects, collection of specimens or experimentation with animals.

## Reference List


Adams, W.M. & Sandbrook, C. (2013) Conservation, evidence and policy. *Oryx*, 47, 329–335.

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.

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

Angrist, J.D. & 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.

Anon. (2019a) Collaboration for Environmental Evidence. http://www.environmentalevidence.org/ [accessed 28 January 2019].

Anon. (2019b) Conservation Evidence. https://www.conservationevidence.com/ [accessed 28 January 2019].

Asquith, N.M. (2016) Watershared: Adaptation, mitigation, watershed protection and economic development in Latin America. Climate & Development Knowledge Network, London.

Babad, E.Y., Inbar, J. & Rosenthal, R. (1982) Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology*, 74, 459–474.

Banerjee, A., Chassang, S. & Snowberg, E. (2016) Decision Theoretic Approaches to Experiment Design and External Validity. NBER Working Paper 22167, Cambridge, MA.

Barton, S. (2000) Which clinical studies provide the best evidence? *BMJ*, 321, 255–256.

Bausell, R.B. (2015) The Design and Conduct of Meaningful Experiments Involving Human Participants: 25 Scientific Principles. Oxford University Press, New York.

Baylis, K., Honey-Rosés, J., Börner, J., Corbera, E., Ezzine-de-Blas, D., Ferraro, P.J., et al. (2016) Mainstreaming Impact Evaluation in Nature Conservation. *Conservation Letters*, 9, 58–64.

Bloom, H.S. (2008) The Core Analytics of Randomized Experiments for Social Research. In *The SAGE Handbook of Social Research Methods* (eds P. Alasuutari, L. Bickman & J. Brannen), pp. 115–133. SAGE Publications Ltd, London.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. & Moore, L. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, 75, 2299–2306.

Börner, J., Baylis, K., Corbera, E., Ezzine-de-Blas, D., Ferraro, P.J., Honey-Rosés, J., et al. (2016) Emerging Evidence on the Effectiveness of Tropical Forest Conservation. *PLOS ONE*, 11, e0159152.

Börner, J., Baylis, K., Corbera, E., Ezzine-de-Blas, D., Honey-Rosés, J., Persson, U.M. & Wunder, S. (2017) The Effectiveness of Payments for Environmental Services. *World Development*, 96, 359–374.

Bottazzi, P., Wiik, E., Crespo, D. & Jones, J.P.G. (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.

BRODY, H. (2012) A critique of clinical equipoise. In *The Ethical Challenges of Human Research* (ed F. Miller), pp. 199–216. Oxford University Press, New York.

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.

BUTSIC, V., LEWIS, D.J., RADELOFF, V.C., BAUMANN, M. & KUEMMERLE, T. (2017) Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, 19, 1–10.

CARTWRIGHT, N. (2010) What are randomised controlled trials good for? *Philosophical Studies*, 147, 59–70.

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.

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, e12391.

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* pp. 269–298. U.S. Government Printing Office, Washington DC.

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

DEATON, A. (2010) Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48, 424–455.

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

DONNELLY, C.A., WOODROFFE, R., COX, D.R., BOURNE, F.J., CHEESEMAN, C.L., CLIFTON-HADLEY, R.S., ET AL. (2005) Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature*, 439, 843–846.

EWERS, R.M., DIDHAM, R.K., FAHRIG, L., FERRAZ, G., HECTOR, A., HOLT, R.D., ET AL. (2011) A large-scale forest fragmentation experiment: The stability of altered forest ecosystems project. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366, 3292-3302.

FERRARO, P.J. & HANAUER, M.M. (2014) Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annual Review of Environment and Resources*, 39, 495–517.

FERRARO, P.J. & PATTANAYAK, S.K. (2006) Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments. *PLoS Biology*, 4, e105.

GILLSON, L., BIGGS, H., SMIT, I.P.J., VIRAH-SAWMY, M. & ROGERS, K. (2019) Finding Common Ground between Adaptive Management and Evidence-Based Approaches to Biodiversity Conservation. *Trends in Ecology & Evolution*, 34, 31–44.

GLENNERSTER, R. & TAKAVARASHA, K. (2013) Running Randomized Evaluations: A Practical Guide. Princeton University Press, Princeton, NJ.

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

HAYNES, L., SERVICE, O., GOLDACRE, B. & TORGERSON, D. (2012) Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. UK Government Cabinet Office Behavioural Insights Team, London.

INDEPENDENT EVALUATION GROUP (2012) World Bank Group Impact Evaluations: Relevance and Effectiveness. World Bank Group, Washington DC.

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, 267–273.

KONTOLEON, A., CONTEH, B., BULTE, E., LIST, J.A., MOKUWA, E., RICHARDS, P., ET AL. (2016) The impact of conditional and unconditional transfers on livelihoods and conservation in Sierra Leone, 3ie Impact Evaluation Report 46. New Delhi.

LEIGH, A. (2018) Randomistas: How Radical Researchers are Changing our World. Yale University Press, New Haven, CT.

LEVITT, S.D. & LIST, J.A. (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.

LIST, J.A. & RASUL, I. (2011) Field Experiments in Labor Economics. In *Handbook of Labor Economics* (eds O. Ashenfelter & D. Card), pp. 104–228. North Holland, Amsterdam.

MCCARTHY, M.A. & POSSINGHAM, H.P. (2007) Active adaptive management for conservation. *Conservation Biology*, 21, 956–963.

MCINTOSH, E.J., CHAPMAN, S., KEARNEY, S.G., WILLIAMS, B., ALTHOR, G., THORN, J.P.R., ET AL. (2018) Absence of evidence for the conservation outcomes of systematic conservation planning around the globe: A systematic map. *Environmental Evidence*, 7, 22.

MEALS, D.W., DRESSING, S.A. & DAVENPORT, T.E. (2010) Lag time in water quality response to best management practices: a review. *Journal of Environmental Quality*, 39, 85–96.

MICHALOPOULOS, C., BLOOM, H.S. & HILL, C.J. (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.

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

PATTANAYAK, S.K. (2009) Rough Guide to Impact Evaluation of Environmental and Development Programs. South Asian Network for Development and Environmental Economics, Kathmandu, Nepal.

PRESSEY, R.L., WEEKS, R. & GURNEY, G.G. (2017) From displacement activities to evidence-informed decisions in conservation. *Biological Conservation*, 212, 337–348.

PULLIN, A.S., KNIGHT, T.M., STONE, D.A. & CHARMAN, K. (2004) Do conservation managers use scientific evidence to support their decision-making? *Biological Conservation*, 119, 245–252.

PYNEGAR, E.L. (2018) The use of Randomised Control Trials in evaluating conservation interventions: the case of Watershared in the Bolivian Andes. Ph.D. Thesis, Bangor University.

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

Rafidimanantsoa, H.P., Poudyal, M., Ramamonjisoa, B.S. & Jones, J.P.G. (2018) Mind the gap: the use of research in protected area management in Madagascar. *Madagascar Conservation and Development*, 13, 15-24.

Rasolofoson, R.A., Ferraro, P.J., Jenkins, C.N. & Jones, J.P.G. (2015) Effectiveness of Community Forest Management at reducing deforestation in Madagascar. *Biological Conservation*, 184, 271–277.

Ravallion, M. (2009) Should the Randomistas Rule? *The Economists' Voice*, 6, 8–12.

Rose, D.C., Sutherland, W.J., Amano, T., González-Varo, J.P., Robertson, R.J., Simmons, B.I., et al. (2018) The major barriers to evidence-informed conservation policy and possible solutions. *Conservation Letters*, 11, e12564.

Rosenthal, R. & Jacobson, L. (1968) Pygmalion in the classroom. *The Urban Review*, 3, 16–20.

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

Rundlöf, M., Andersson, G.K.S., Bommarco, R., Fries, I., Hederström, V., Herbertsson, L., et al. (2015) Seed coating with a neonicotinoid insecticide negatively affects wild bees. *Nature*, 521, 77–80.

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

Segan, D.B., Bottrill, M.C., Baxter, P.W.J. & Possingham, H.P. (2011) Using Conservation Evidence to Guide Management. *Conservation Biology*, 25, 200–202.

Sutherland, W.J. & Wordley, C.F.R. (2017) Evidence complacency hampers conservation. *Nature Ecology & Evolution*, 1, 1215–1216.

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

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

Walsh, J.C., Dicks, L. V. & Sutherland, W.J. (2015) The effect of scientific evidence on conservation practitioners' management decisions. *Conservation Biology*, 29, 88–98.

White, H. (2013a) The Use of Mixed Methods in Randomized Control Trials. *New Directions for Evaluation*, 2013, 61–73.

White, H. (2013b) An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness*, 5, 30–49.

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

WHO (2013) WHO Evaluation Practice Handbook. World Health Organization, Geneva.

Wiik, E., D'Annunzio, R., Pynegar, E.L., Crespo, D., Asquith, N.M. & Jones, J.P.G. (2019) Can Payments for Environmental Services reduce deforestation? Results from a Randomized Control Trial experiment in the Río Grande catchment of Bolivia. *Conservation Science and Practice*, e8.

Woodcock, B.A., Bullock, J.M., Shore, R.F., Heard, M.S., Pereira, M.G., Redhead, J., et al. (2017) Country-specific effects of neonicotinoid pesticides on honey bees and wild bees. *Science*, 356, 1393–1395.

## Tables

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 Hawthorne effect discussed in Levitt & List (2011).

| Effect name | Description/Explanation | Effect on outcome in treatment group | Effect on outcome in control group | Effect on estimated effect size of intervention |
|---|---|---|---|---|
| 'Hawthorne 1' | Evaluators being seen to observe participants causes participants to increase effort. | Increases | Increases | Unknown |
| 'Hawthorne 2' | Modifications made to the intervention itself during the course of the experiment cause participants to increase effort. | None / Increases | None | None / Increases |
| 'Hawthorne 3' | Experimental participants tend to meet what they believe to be experimenters' expectations. This may derive from increased effort in treatment units (the *Pygmalion effect*; Rosenthal & Jacobson, 1968) and/or decreased effort in control units (the *golem effect*; Babad et al., 1982). Treatment-group interviewees also tend to give answers they believe evaluators wish to hear (*experimenter demand*; Levitt & List, 2011). | Increases | None / Decreases | Increases |
| Rational effort | Experimental participants decide how much effort to expend on implementing an intervention based upon their own expectations of the intervention's effectiveness; this closely parallels the *Galatea effect* (Babad et al., 1982). | Increases | None / Decreases | Increases |
| 'John Henry' | Individuals in the control group increase effort in an attempt to compete with the intervention group (Saretsky, 1972; see also Bausell, 2015). | None | None / Increases | None / Decreases |

## Figures

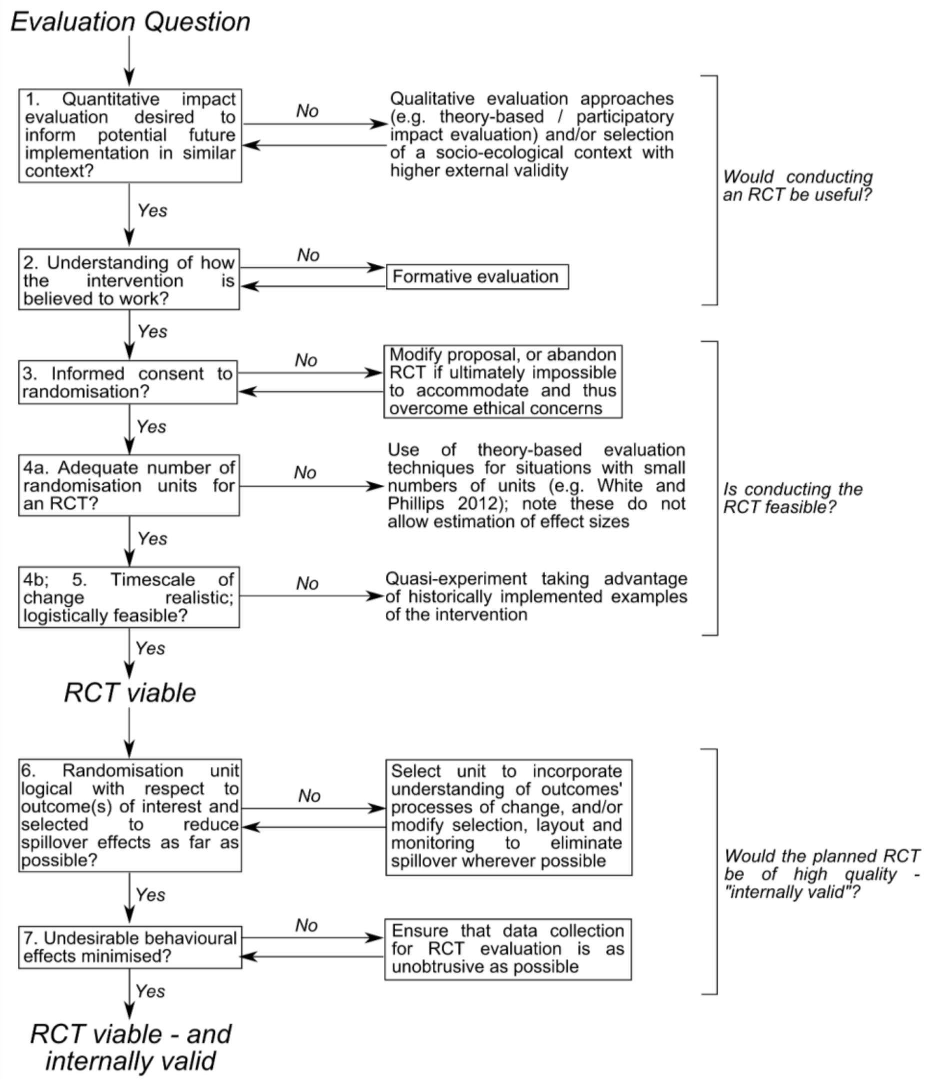


Figure 1. Summary of suggested decision-making process for evaluators to decide if an RCT evaluation of their conservation intervention would useful, feasible, and of high quality.

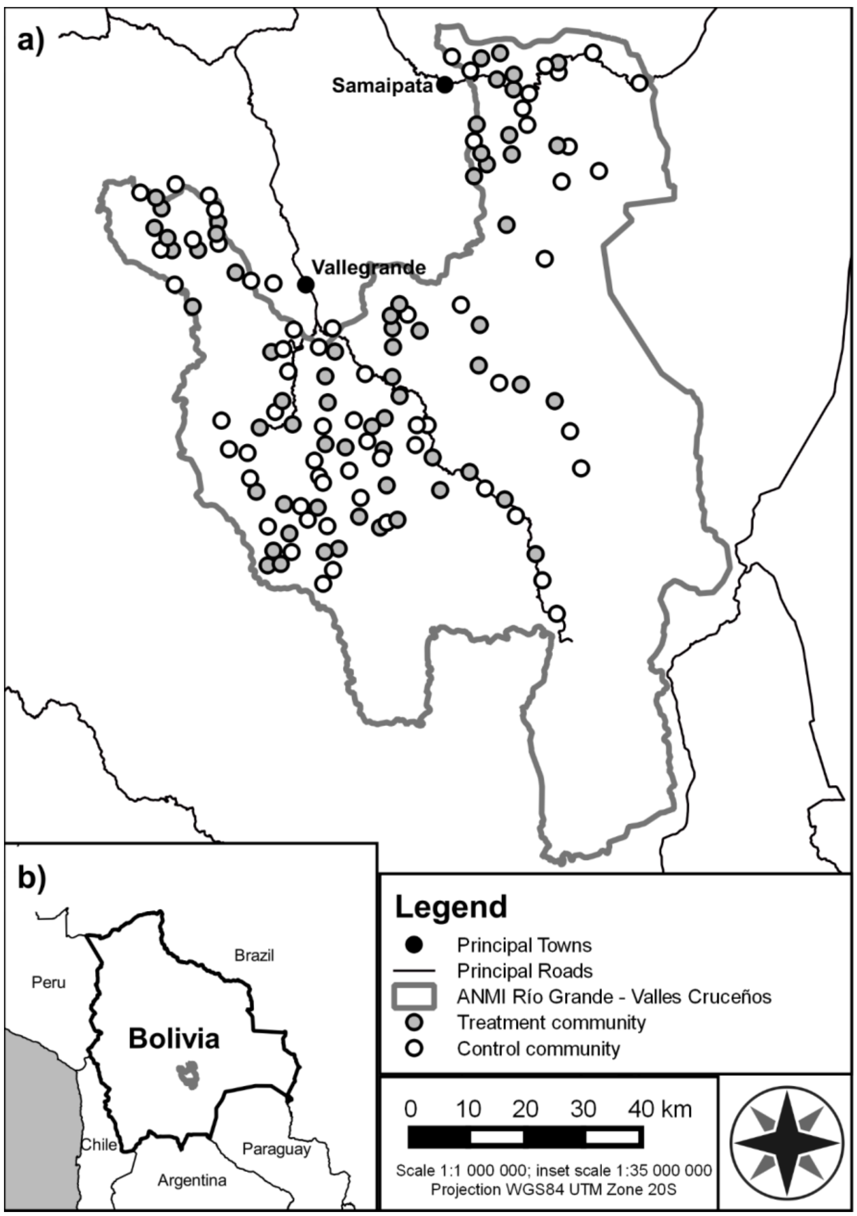


Figure 2. a) Locations of the 65 treatment and 64 control communities included in the RCT evaluating the impact of the Watershared incentive-based conservation intervention by Fundación Natura Bolivia

(‘Natura’) in the Bolivian Andes. b) Location of the RCT (the ANMI Río Grande – Valles Cruceños protected area) within Bolivia.