Review

What role should randomized control trials play in providing the evidence base for conservation?

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

Abstract The effectiveness of many widely used conservation interventions is poorly understood because of a lack of high-quality impact evaluations. Randomized control trials (RCTs), in which experimental units are randomly allocated to treatment or control groups, offer an intuitive way to calculate the impact of an intervention by 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 how 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 few concluded RCTs of a large-scale conservation intervention: an incentive-based conservation programme in the Bolivian Andes. We argue that conservation should aim to avoid a rerun of the polarized debate surrounding the use of RCTs in other fields. Randomized control trials will not be feasible or appropriate in many circumstances, but if used carefully they can be useful and could become a more widely used tool for the evaluation of conservation impact.

Keywords Counterfactual, effectiveness, evidence, impact evaluation, randomization, randomized control trials, RCTs

Edwin L. Pynegar (Corresponding author, orcid.org/0000-0001-5975-696X), James M. Gibbons and Julia P. G. Jones College of Environmental Sciences and Engineering, Bangor University, Bangor, Gwynedd, LL57 2UW, UK
E-mail edwin.pynegar@gmail.com

Nigel M. Asquith* Harvard Forest, Petersham, USA

*Also at: Sustainability Science Program, Harvard Kennedy School, Cambridge, USA

Received 20 June 2018. Revision requested 6 November 2018.
Accepted 13 February 2019. First published online 24 October 2019.

Introduction

It is widely recognized that conservation decisions should be informed by evidence (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). Although this is at least partly a result of 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 (Rafidimanantsoa et al., 2018; Rose et al., 2018), there are limitations in the evidence available 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 systematize the existing evidence, such as The Collaboration for Environmental Evidence (2019) and Conservation Evidence (2019).

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 an 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 an 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) that lack the intervention. This is increasingly used for conservation impact evaluations, such as determining the impact of establishment of a national park (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 the results (Michalopoulos et al., 2004; Glennerster & Takavarasha, 2013). historically this problem led many in development economics to question the usefulness of such quasi-experiments (Angrist & Pischke, 2010). Each kind of quasi-experiment has associated assumptions that, if not met, affect the validity of the evaluation result (Glennerster & Takavarasha, 2013).

Randomized control trials (RCTs; also known as randomized 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 that will receive a particular intervention (the treatment group), and those that will not (the control group), there should be no systematic differences between groups (White, 2013a). 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 to sceptical audiences than other impact evaluation methods (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 may or may not work (Glennerster & Takavarasha, 2013). Randomized control trials 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 the efficacy of treatments (Barton, 2000). They are also widely used in agriculture, education, social policy (Bloom, 2008), labour economics (List & Rasul, 2011) and increasingly in development economics (Ravallion, 2009; Banerjee et al., 2016; Deaton & Cartwright, 2018; Leigh, 2018). The governments of both the UK and the USA have strongly supported the use of RCTs in evaluating policy effectiveness (Haynes et al., 2012; Council of Economic Advisers, 2014). The U.S. 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 (Cartwright, 2010) and practical (Deaton, 2010; Deaton & Cartwright, 2018) critiques of RCTs. The statistical basis of randomized analyses is also not necessarily simple. Randomization can only be guaranteed to lead to complete balance between treatment and control groups with extremely large samples (Bloom, 2008), although 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 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 polarized debate that the spread of RCTs 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; Ferraro & Hanauer, 2014; Samii et al., 2014; Baylis et al., 2016; Börner et al., 2016, 2017; Curzon & Kontoleon, 2016). As many kinds of conservation programmes 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+ programmes, and certification programmes; we term these socio-ecological interventions), there are 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 programme in Bolivia known as Watershared, described here; a payment programme 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 programme to reduce wild meat consumption in the Brazilian Amazon through social marketing and incentivising consumption of chicken (Chaves et al., 2018). We expect that evaluation with RCTs will become more widespread in conservation.

Here 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 text is mirrored by a checklist (Fig. 1) that can be used to assess the suitability of an RCT in a given context. We illustrate these points with the RCT evaluating the Watershared incentive-based conservation programme in the Bolivian Andes. This programme, 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).

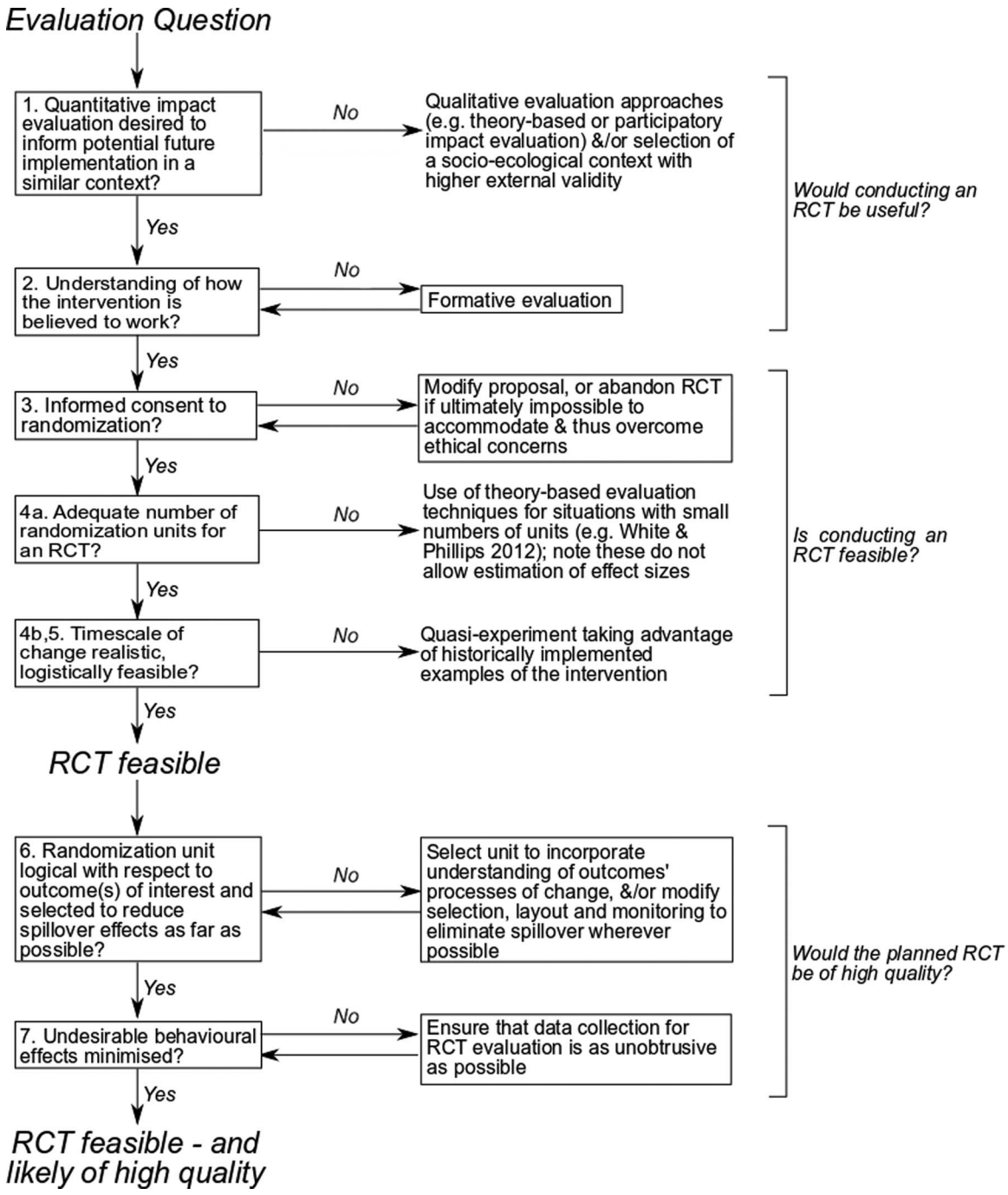

FIG. 1 Summary of suggested decision-making process to help decide whether a randomized control trial (RCT) evaluation of a conservation intervention would be useful, feasible and of high quality. Items in the right-hand column without a box represent end-states of the decision-making process (i.e. an RCT is probably not appropriate and the researcher should consider using an alternative evaluation method).

Under what circumstances could an RCT evaluation be useful?

When quantitative evaluation of an intervention's impact is required

Randomized control trials 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 the theory of change may be more suitable where such quantitative estimates are not needed or where the intervention can only be implemented in a 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 et al., 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 use of RCTs in development economics (e.g. Cartwright, 2010; Deaton, 2010). Advocates for RCTs accept such critiques as partially valid (e.g. White, 2013a) and acknowledge that RCTs should be considered to provide knowledge that is complementary to, not incompatible with, other approaches. Firstly, qualitative studies can be conducted alongside an RCT to examine processes of change; most evaluators who advocate RCTs also recognize that combining quantitative and qualitative approaches is likely to be most informative (e.g. White, 2013b). Secondly, researchers can use covariates to explore which contextual features affect outcomes of interest, to look for those features in future implementation of the intervention (although to avoid data dredging, hypotheses and analysis plans should ideally 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 be designed to answer more complex contextual questions through trials with multiple treatment groups or other modifications (Bonell et al., 2012). Thirdly, evaluators may conduct RCTs of the same kind of intervention in different socio-ecological contexts (White, 2013a), which increases the generalizability of results. Although this is challenging because of the spatial and temporal scale of RCTs used to evaluate socio-ecological interventions, researchers have undertaken a number of RCTs of incentive-based conservation programmes (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 used to evaluate the Bolivian Watershared programme, 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. 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) and to compare outcomes in control and treatment communities (Pynegar et al., 2018; Wiik et al., 2019).

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 because once begun, 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 they have an intervention whose functioning is reasonably well understood (Pattanayak, 2009; Cartwright, 2010).

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 that 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 programme's effectiveness and the evaluation's usefulness. For example, uptake of the agreements was 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 programme on water quality at the landscape scale could 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

Randomization involves withholding the intervention from the control group, so the decision to randomize is not a morally neutral one. An ethical principle in medical RCTs is that to justify a randomized 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 will often 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 it is also important to know for policymakers how well an intervention works and how cost-effective it is (White, 2013a). 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 handled sensitively (White, 2013a).

Another principle of research ethics is 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 could be overcome by randomizing 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 could 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 randomization and individual landowners could choose to sign an agreement or not. Although 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, hampering the subsequent evaluation (Wiik et al., 2019).

Spatial and temporal scale

Larger numbers of randomization 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 are a mainstay of applied ecology. However, increases in the scale of the intervention will make RCT implementation more challenging. Interventions implemented at a large scale will probably have few randomization 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 randomization units are also likely to increase costs and logistical difficulties. However, 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 low, however, RCTs will not be appropriate and evaluations based upon analysing expected theories of change may be more advisable (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 because of long species life cycles or the slow and stochastic nature of 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 probably consists 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 randomization units, each consisting of 2–185 households). Following baseline data collection in 2010, the intervention was first offered in 2011 and endline data was collected in 2015–2016. Effects on water quality were expected to be observable over this timescale as cattle exclusion can result in decreases in waterborne bacterial concentration in < 1 year (Meals et al., 2010). However, there was no impact of the intervention on water quality at the landscape scale (Pynegar et al., 2018), potentially because of time lags; nor did the programme significantly reduce deforestation rates (Wiik et al., 2019). A potential explanation is that impacts may take longer to materialize as they could depend on the development of alternative livelihoods introduced as part of the programme.

Available resources

Randomized control trials require substantial human, financial and organizational resources for their design, implementation, monitoring and evaluation. These resources are above the additional cost of monitoring in control units, because design, planning, and subsequent analysis and interpretation require substantial effort and knowledge. USAID advises that a minimum of 3% of a project or programme's budget be allocated to external evaluation (USAID, 2016), and 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 expensive (Curzon & Kontoleon, 2016).

Collaborations between researchers (with independent funding) and practitioners (with a part of their programme 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 organizations, and the additional costs of the RCT were covered by 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 (2019), the Abdul Latif Jameel Poverty Action Lab (2019) and the International Initiative for Impact Evaluation (2019).

What factors affect the quality of an RCT evaluation?

Potential for spillover, and how selection of randomization 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; in conservation they could also potentially include fields, farms, habitat patches, protected areas, or other units. 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 the functioning of interventions. 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 randomization 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 programme 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 was to be implemented in all areas rather than solely in 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: evaluation of badger culling in south-west England (Donnelly et al., 2005).

Spillover is particularly likely if the randomization unit and the natural unit of the intended ecological process of change are incongruent, meaning the intervention would inevitably be implemented in areas that would affect outcomes in control units. Therefore, consideration of spatial relationships between units, and of the relationship between randomization units and the outcomes' process of change, is critical. For example the anti-poaching programme described above could instead use closed groups or populations of the target species as the randomization unit, with the programme then implemented in communities covering the range of each treatment group. Spillover may also be reduced by selecting indicators (and/or sites to monitor) that 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 that has a small range or ensuring that a control area's monitoring site is not directly downstream of that of a treatment area in an RCT of a payments for watershed services programme).

In the RCT of Watershared, it proved difficult to select a randomization unit that was politically feasible and worked for all outcomes of interest. Natura used community as the randomization unit, so community boundaries had to be defined but these did not always align well with the watersheds supplying the communities' water sources. Although few water quality monitoring sites were directly downstream of another, land under agreements in one community were in some cases 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 randomization unit worked well for the deforestation analysis. Communities have definable boundaries (although see Wiik et al., 2019) and offering the programme by community was most practical logistically. A smaller unit would have presented issues of perceived fairness as it would have been difficult to offer Watershared agreements to some members of communities and not to others. The RCT of Jayachandran et al. (2017) also selected community as the randomization 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 that 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 (Rundlöf et al., 2015). However, it is harder to carry out double-blind trials of socio-ecological interventions, as the intervention's consequences

TABLE 1 Consequences of behavioural effects when compared with results obtained in a hypothetical double-blind randomized control trial. 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) &/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

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 by selecting units whose individuals infrequently interact with each other. Evaluators of Watershared believed that members of control communities could decide to protect watercourses themselves after seeing successful results elsewhere (which would be encouraging for the NGO, suggesting local support for the intervention, but that 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 $>$ 1,500 household surveys (Pynegar, 2018).

The second issue with lack of blinding is that randomization is intended to ensure that treatment and control groups are not systematically different immediately after randomization. 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 as a result of differences in effort between treatment and control groups (summarized 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 because of effort expended by the implementers (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 (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 the use of an intervention 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 nevertheless 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.

As with any evaluation method, RCTs 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 will also inevitably remain some level of subjectivity as to whether findings from an RCT are applicable with confidence to a different location or context. 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 but avoid the polarization and controversy surrounding them. Randomized control trials could then make a substantial contribution towards the evaluation of conservation impact.

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). NMA 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 Programme. We thank our colleagues and collaborators at Fundación Natura Bolivia, particularly María Teresa Vargas and Tito Vidaurre, for valued discussion, Jörn Scharlemann for helpful comments, and two anonymous reviewers for their valuable critiques.

Author contributions Literature review: ELP; writing: all authors.

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

Ethical standards This research abided by the *Oryx* guidelines on ethical standards.

References

Abdul Latif Jameel Poverty Action Lab (2019) Http://www.povertyactionlab.org [accessed 11 August 2019].

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.

Asquith, N.M. (2016) *Watershared: Adaptation, Mitigation, Watershed Protection and Economic Development in Latin America*. Climate & Development Knowledge Network, London, UK.

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, USA.

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, USA.

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, UK.

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, USA.

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.

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

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, USA.

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, 3,292–3,302.

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, USA.

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, USA.

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

Innovations for Poverty Action (2019) Http://www.poverty-action.org [accessed 11 August 2019].

International Initiative for Impact Evaluation (2019) Http://www.3ieimpact.org [accessed 11 August 2019].

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, India.

Leigh, A. (2018) *Randomistas: How Radical Researchers Are Changing Our World*. Yale University Press, New Haven, USA.

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, Netherlands.

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*. PhD thesis, Bangor University, Bangor, UK.

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, USA.

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*, 2014, 11.

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.

The Collaboration for Environmental Evidence (2019) Http://www.environmentalevidence.org [accessed 28 January 2019].

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

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

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

White, H. (2013b) The use of mixed methods in randomized control trials. *New Directions for Evaluation*, 2013, 61–73.

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, India.

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

Wiik, E., D'Annunzio, R., Pynegar, E.L., Crespo, D., Asquith, N.M. & Jones, J.P.G. (2019) Experimental evaluation of the impact of a payment for environmental services program on deforestation. *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.