Received: 4 October 2019 | Revised: 22 February 2020 | Accepted: 12 April 2020

DOI: 10.1111/conl.12726

Conservation Letters
A journal of the Society for Conservation Biology
Open Access
WILEY

REVIEW

# Improving scientific rigour in conservation evaluations and a plea deal for transparency on potential biases

**Jonas Josefsson**[1] | **Matthew Hiron**[1] | **Debora Arlt**[1] | **Alistair G. Auffret**[1] | **Åke Berg**[2] | **Mathieu Chevalier**[1,3] | **Anders Glimskär**[1] | **Göran Hartman**[1] | **Ineta Kačergytė**[1] | **Julian Klein**[1] | **Jonas Knape**[1] | **Ane T. Laugen**[1,4] | **Matthew Low**[1] | **Matthieu Paquet**[1] | **Marianne Pasanen-Mortensen**[1,5] | **Zuzanna M. Rosin**[1,6] | **Diana Rubene**[1,7] | **Michał Żmihorski**[1,8] | **Tomas Pärt**[1]

[1]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[2]Swedish Biodiversity Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden

[3]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[4]Department of Natural Sciences, Centre for Coastal Research, University of Agder, Kristiansand, Norway

[5]Department of Zoology, Stockholm University, Stockholm, Sweden

[6]Department of Cell Biology, Institute of Experimental Biology, Faculty of Biology, Adam Mickiewicz University, Umultowska, Poznań, Poland

[7]Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[8]Mammal Research Institute, Polish Academy of Sciences, Białowieża, Poland

**Correspondence**
Jonas Josefsson & Tomas Pärt, Department of Ecology, Swedish University of Agricultural Sciences, Box 7044, 75007 Uppsala, Sweden.
Email: jonas.josefsson@slu.se, tomas.part@slu.se

**Abstract**

The delivery of rigorous and unbiased evidence on the effects of interventions lay at the heart of the scientific method. Here we examine scientific papers evaluating agri-environment schemes, the principal instrument to mitigate farmland biodiversity declines worldwide. Despite previous warnings about rudimentary study designs in this field, we found that the majority of studies published between 2008 and 2017 still lack robust study designs to strictly evaluate intervention effects. Potential sources of bias that arise from the correlative nature are rarely mentioned, and results are still promoted by using a causal language. This lack of robust study designs likely results from poor integration of research and policy, while the erroneous use of causal language and an unwillingness to discuss bias may stem from publication pressures. We conclude that scientific reporting and discussion of study limitations in intervention research must improve and propose some practices toward this goal.

**KEYWORDS**

agri-environment scheme, before after control impact, biodiversity | causal language, evaluation of conservation interventions, meta-analysis, organic farming, study design, systematic review

## 1 | INTRODUCTION

Biodiversity loss has direct, tangible effects on ecosystem functioning and human well-being (Cardinale et al., 2012). There is a growing body of research in conservation ecology, which should provide the solid ground needed for meta-analysis and synthesis, leading to effective evidence-based environmental management (Pullin & Knight 2001; 2009;

Sutherland, Pullin, Dolman, & Knight, 2004). Naturally, the value of synthesis research relies on the quality of the underlying evidence. In conservation research, the scarcity of experimental and longitudinal studies (Ferraro, 2009; Ferraro &Pattanayak, 2006) translates into correlative and bias-prone evidence, which is then being fed into systematic reviews and syntheses (Haddaway & Bilotta, 2016). With this in mind, it is important that systematic reviews provide a critical appraisal of internal (bias susceptibility) and external (study relevance) validity of included studies (Collaboration for Environmental Evidence, 2013). However, failing to report limitations complicates such assessments.

Concerns about the lack of disclosure of bias and other limitations in original studies have been expressed previously in the fields of epidemiology and public health sciences (where evidence-synthesis methods originated), with calls for transparent and systematic reporting of study limitations (Puhan etal., 2012; ter Riet etal., 2013). The use of observational methodologies also constrains causal inferences, but misuse of causal claims is still common across disciplines (Cofield, Corona, & Allison, 2010; Robinson, Levin, Thomas, Pituch, & Vaughn, 2007). In this paper, we use the example of agri-environmental schemes (AES) to demonstrate that these problems are also widespread in environmental sciences. AES are the primary policy instruments used to safeguard biodiversity and ecosystem services in agricultural landscapes worldwide, including North America (Stubbs, 2013), Australia (Burns, Zammit, Attwood, & Lindenmayer, 2016), Africa (Kehinde & Samways, 2014), and Asia (Nomura etal., 2013). In Europe, more than €20 billion was spent on such schemes between 2007 and 2013 (Science for Environment Policy, 2017). Considering the importance of the matter, and the high costs involved, well-designed evaluations are central to understand the mechanisms and impacts of different conservation interventions under diverse agricultural contexts.

What are the caveats that we, as researchers in environmental sciences, must acknowledge and discuss? First, *nonrandom patterns of implementation of conservation programs preclude effective evaluation of their success*. This situation arises from large-scale conservation programs typically being implemented before dedicated evaluations are outlined. While seldom considered, this is critical when evaluating interventions as it precludes the use of randomized experimental designs and sampling before and after an intervention. Use of designs that allow stronger causal inference, including randomized controlled trials or observational before-after-control-impact (BACI) designs (Figure 1), is therefore often not possible. Instead, researchers are constrained to adopt weaker observational designs (Christie et al., 2019). These study designs, which include control-impact (CI) studies, are highly susceptible to bias from the selection of intervention areas, where selection probability correlates with conditions that themselves affect biodiversity baselines and responses (Ferraro, 2009; Ferraro &Pattanayak, 2006). For example, a conservation action is more likely to be implemented at a location where it is expected to work or where original biodiversity is high. An example of this is the targeting of biodiversity-rich areas for protection and management in conservation planning (Brooks et al., 2006; Eken et al., 2004; Groves et al., 2002; Myers, Mittermeier, Mittermeier, da Fonseca, & Kent, 2000). In agricultural and forested areas, where participation in environmental schemes is often encouraged by financial compensation, this effect may be less obvious. The landowners or managers that are more likely to participate in such incentive schemes may differ from nonparticipants across key variables that, in themselves, are important drivers of biodiversity patterns, such as management intensity, soil fertility, landscape complexity, and microclimate (Gabriel et al., 2009). Even when methods are used to adjust for any such known differences across sites, important and *unknown* confounding effects may still be left unaccounted for (Little & Rubin, 2000). These features of conservation programs mean that impact assessments using observational methods are at best uncertain, at worst apparently flawed, especially when there are no data recorded before the intervention occurred.

The second caveat is *the potential misuse of causal language in observational studies*. Observational CI studies produce potentially biased data in terms of what is driving observed effects, and where the initial selection of "impact" sites is a central problem for making causal inferences (Elwert & Winship, 2014). This begs the question: Is it right to infer causal effects of interventions, whether in primary studies or in reviews when the underlying data typically is of an observational and bias-prone nature? As mentioned in any book on study design and scientific methods, observational study designs are generally restricted in terms of their capacity for causal inference (Underwood, 1997). The main problem of implying causation from correlative observations is that it may divert attention from the real reasons for any observed effect, promoting false confidence in the drivers of the observed pattern.

More than a decade after the widespread implementation of AES across Europe, Kleijn and Sutherland highlighted the need for improved study design in conservation evaluations in their seminal review on the effectiveness of AES for the conservation of biodiversity published in 2003 (Kleijn & Sutherland, 2003). In a comprehensive search of the scientific literature, they found that inadequate research designs prevented a reliable assessment of measures that had been implemented. Since then, the number of scientific evaluations of AES has grown considerably (see Ansell, Freundenberger, Munro, & Gibbons, 2016) and includes several reviews and meta-analyses. Given the vast extent of these policy instruments—in terms of geographic spread, financial investment, and public interest—quite some trust is placed on how we scientists

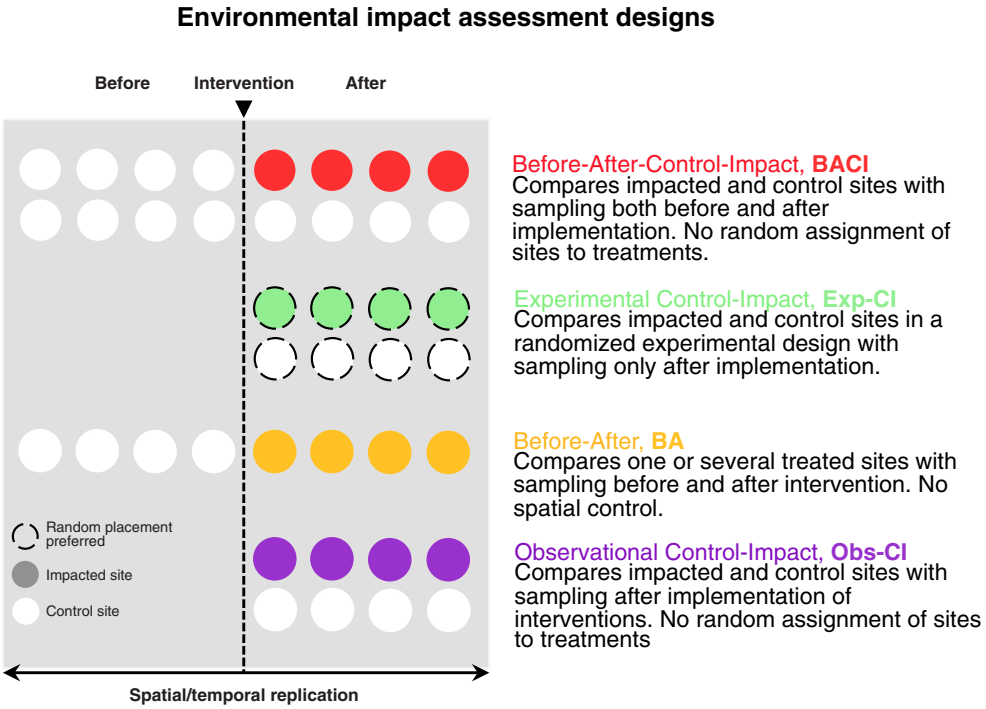


**FIGURE 1** Study designs used to valuate effects of conservation interventions and ecosystem services

evaluate these interventions. Allowing a grace period of 5 years for new studies to be carried out since their publication, we examined scientific evaluations of the effects of AES on biodiversity published over the following 10 years 2008–2017 to investigate (i) if more recent evaluations have improved in terms of study design and the extent to which potential limitations associated with selection bias are acknowledged and (ii) the prevalence of causal statements, particularly in studies with observational data. As the benefits of organic farming are regularly debated in scholarly journals (Balmford et al., 2019; Eyhorn et al., 2019) and in news media (Reganold, 2016; Savage, 2015), we were specifically interested in this policy option and therefore chose to separate studies into evaluations of organic farming and other AES, respectively. Such interventions are wide-ranging, but generally include support for extensive farming practices such as low-intensity grazing and management of landscape features of high natural or historical value. While we focus on AES, these concerns are common to environmental policies and their evaluation in other human-impacted environments, including forests (França et al., 2016; Wikberg et al., 2009) and marine systems (de Loma et al., 2008; Osenberg, Shima, Miller, & Stier, 2011).

## 2 | METHODS

We searched for original research papers published from 2008 to 2017 in peer-reviewed scientific journals using a predefined search and screening protocol (see the Supporting Information for details). From the 215 resulting studies, we extracted information on (1) intervention type (organic farming or other AES), (2) study design (observational control-impact [obs-CI], before-after [BA], experimental, and randomized control-impact [exp-CI], BACI; see Figure 1), (3) acknowledgement of, and accounting for the potential for baseline biases in biodiversity between impact and control sites (paired design, use of covariates, or other types of reducing baseline biases, Supplementary Appendix), and (4) causal terminology used by authors to describe results. Although the term BACI is reserved for observational studies (Underwood, 1992), in this category we included also two experimental studies that used before-after data to highlight the limited occurrence of collecting data before interventions. We also searched literature syntheses published during the same time period and we collected similar data as for the original studies ($n = 22$ reviews). For details about data coding see the Supporting Information. Concerning "causal statement coding" we searched for sentences containing definitive causal language or hedged versions (e.g., ("can", "may") in the title, abstract, and discussion sections. Similarly we searched the abstract, methods, and discussion sections to determine the rate that studies reported on study limitations relating to study design and implications for internal validity (for details, see the Supporting Information).

The full set of coded papers and the codes is available online (Supplementary Appendix).

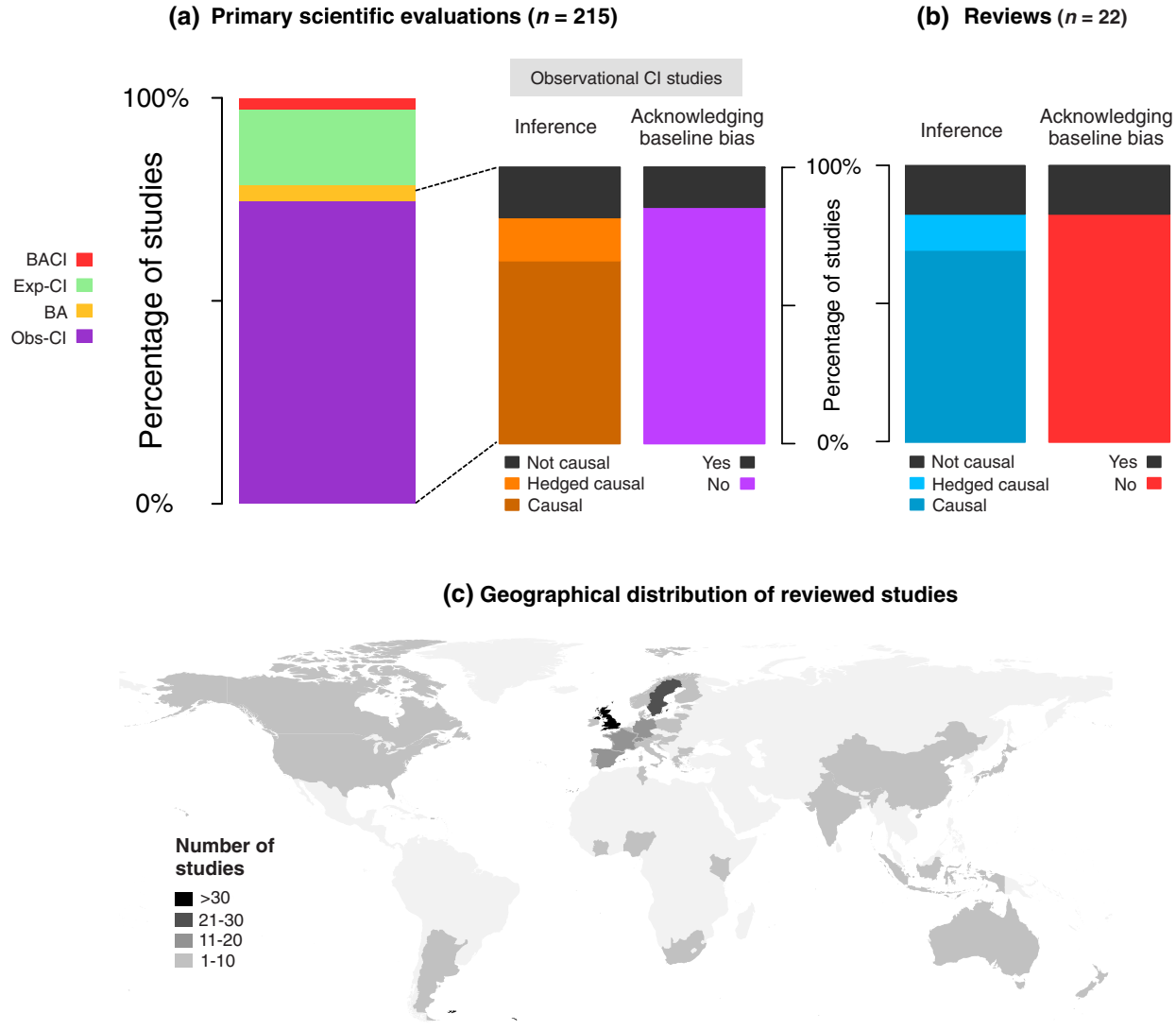


**FIGURE 2** Scientific reporting of results and bias-related limitations in (a) primary evaluations and (b) reviews of the value of AES, including organic farming, for biodiversity. In observational control-impact studies (Obs-CI), incorrect causal inference was prevailing and study limitations seldom discussed. (c) Studies were distributed globally, but with a concentration of studies to Europe, North America, and Asia. Study designs also included before-after (BA), experimental control-impact (Exp-CI), and before-after-control-impact (BACI) design

## 3 | RESULTS

Of the 215 reviewed studies, 123 evaluated the biodiversity effects of organic farming, while 92 described the effects of other AES measures. A majority (74%) of the evaluations used observational control-impact designs (80% and 67% of organic farming and AES studies, respectively), while only 19% used an experimental control-impact design and 3% a BACI design (2% on observational data, 1% on experimental before-after data) (Figure 2a; Table S1 in the Supporting Information). Of the observational CI studies, only 14% explicitly mentioned the risk of unaccounted initial bias in biodiversity between control and impact sites either in the abstract, methods, or discussion sections (6% among organic farming and 27% among other AES studies; Table S2 in the Supporting Information). On the other hand, many observational CI studies did at least use a paired design (48% of the Obs-CI studies; cf. 16% in Kleijn & Sutherland 2003), and/or included covariates in their statistical models (68%) to account for possible effects of landscape and other environmental variables on local biodiversity (Table S3 in the Supporting Information). Three additional Obs-CI studies mentioned that the selection of sites was made to keep environmental variables a similar as possible between control and impact plots. This gives a grand total of 84% of the studies (i.e., combining all possible bias reduction strategies) that potentially reduced or accounted for biases in initial conditions between control and impact sites. Still, even when bias reduction methods are used, baseline differences may still exist, but this was only acknowledged in 16% of the studies.

Importantly, despite the correlative nature of the data in the Obs-CI studies, definitive (i.e., without hedging) causal wording was common (66%). Hedged causal statements, using words such as "may," "appears to," and "indicates" to soften causal terminology, was used in another 16% of the studies (Figure 2a; Table S4 in the Supporting Information). Here, the use of definitive causal wording was highest for studies having both a paired design and model covariates (77% out of 53 studies), and lowest for those studies not accounting for any possible baseline bias (52% out of 25 studies). Last, studies covered all continents (except Antarctica), but there was a dominance of European studies (Figure 2c). Short study lengths were common (64% of the studies were only 1-year in duration; Figure S1 in the Supporting Information).

Limitations in study design and the dominance of causal language in AES studies also spilled over into reviews and meta-analyses. Only three of 22 reviews (14%; Figure 2b and Table S5 in the Supporting Information) published between 2008 and 2017 mentioned selection bias as a potential source of uncertainty in the interpretation of effects. As these reviews largely cite the same publications as here or in Kleijn and Sutherland (2003), we know that they were generally dominated by observational studies. Still, causal language was highly prevalent also in reviews when discussing any general biodiversity effects of organic farming and AES (65%, 82% including hedged statements; Figure 2b). While some of the reviews mentioned the utility of paired designs when contrasting impacted to control sites or including covariates in analyses, these approaches were generally not discussed in relation to the risk of selection bias but were mentioned in relation to the investigation of landscape dependency of effects.

## 4 | DISCUSSION

Using the example of AES, our study clearly shows that impact evaluations mostly use bias-prone correlative study designs, while simultaneously failing to fully acknowledge this potential source of bias and erroneously using causal language to convey study findings. It is therefore clear that problems still remain in terms of study design, and that calls from the scientific community for the integration of impact evaluation into environmental policies have not materialised (see, e.g., Baylis et al., 2015; Ferraro, 2009; Fisher et al., 2013). A major obstacle for the development of robust evaluation studies is the lack of researcher influence in the design and implementation stages of conservation interventions (Margoluis, Stem, Salafsky, & Brown, 2009). We are aware that the execution of randomly distributed treatment and control sites is difficult considering logistic constraints and the limited funds available for conservation. It may also be untenable, as it would reduce the delivery of direct common goods, as funding would be needed to pay for randomly assigned controls that deliver no clear benefit. Whether it is more costly in the long-term to fund large-scale experiments evaluating the effectiveness of a full range of AES under different contexts that may have few direct benefits for biodiversity, or on the other hand, to implement poorly evaluated and thus possibly ineffective interventions is, however, debatable.

What are the ways forward to circumvent or solve these problems and deliver scientifically sound impact evaluations? Recent initiatives of collaborative networks including policymakers, farmers, and researchers (Berthet et al., 2018) could open up for an integration of evaluation design in the implementation process. Although the problems of self-selection (vs. randomized selection) may remain, such studies at least can be designed to collect before-after data. Another route to improve evaluation designs where an experimental approach is not feasible is to combine before-after data on impact sites with data on background trends collected from national monitoring schemes and citizen science data (i.e., a BACI design; Underwood, 1992). Including original differences in biodiversity between control and impact sites can then be used to detect and categorize the effects of an intervention even when it is hidden by a general negative trend in focal species at regional scales (i.e., at scales larger than covered by the study; Bull, Gordon, Law, Suttle, & Milner-Gulland, 2014), or by original differences in biodiversity (Chevalier, Russell, & Knape, 2019). In Box 1, we outline three scenarios of improving evaluations of conservation actions in the future.

Short of adopting these or other more-or-less causally valid study designs the scientific community, as well as other users of conservation research, would undoubtedly benefit from an open discussion of the limitations to current evaluation methodologies. Worryingly, our findings suggest that authors are generally either unaware of the limitations related to observational approaches, or that they are unwilling to discuss them. Although the use of pair-matching methods or using covariates for reducing bias could in part explain why the explicit acknowledgment of selection bias is poor, it does not support the erroneous use of casual language. It has been suggested that competition among researchers and journals for high impact publications may foster a culture to neglect inherent and fundamental flaws related to study design, or to falsely make causal claims, in order to increase the seeming significance of research findings (Cofield, Corona, & Allison, 2010; Lipton & Ødegaard, 2005; Puhan et al., 2012; Robinson et al., 2007). This is something that many of us have, at one time or another, probably been guilty of. A culture to let study design limitations go by unremarked may also be fostered at the interface between applied sciences and policy when policymakers provide funding and expect clear answers to research questions. Similarly, editors of applied journals may suggest authors to provide clear directives to practitioners (Robinson et al., 2007; but see Cofield et al.,

**Box 1. IMPROVING EVALUATIONS OF AES CONSERVATION ACTIONS**

We use the case of organic farming to envision a way toward better evaluations of AES effects on biodiversity. We outline three potential scenarios, starting with the most robust.

Implementation of organic farming is usually administered at the national level, with governmental funding supporting the conversion from conventional to organic farming. When farmers apply for financial help to convert, we suggest this should be linked to a governmentally funded before-after (BA) inventory of biodiversity (or other target of interest) at converting farms, and, preferably, at a nearby and otherwise similar conventional farm. Selection of farms for the BA-evaluation should be made in close cooperation between the responsible authorities and researchers.

All scenarios require tight links between policy makers and practitioners, researchers, and national environmental protection agencies for implementation of inventories.

**In scenario 1**: BA evaluation among farms that apply to convert to organic farming. Random selection of some farms as "organic" and others as "control." The "organic" farms proceed with the conversion process, while the "control" farms stay conventional for a limited period. Baseline biodiversity data will be gathered before the conversion process, and farms will be resurveyed after a number of years. After the second round of surveys, the control farms can proceed with conversion to organic farming. Scenario 1 ensures an experimental design that minimizes potential biases of self-selection. All applicant farms get similar subsidies for their farm, that is those decided to initially remain conventional will get reimbursed for their delay to convert to organic farming.

**In scenario 2**: BA evaluation among farms that apply to convert to organic farming and at selected existing conventional farms. All selected farms will be subjected to BA inventories at the same time points. This design does not preclude possible biases due to self-selection of organic farming practices, but potential original differences between organic and conventional farms can be handled within a BACI framework (Underwood, 1992) to evaluate the effect at impact sites (see Chevalier et al., 2019).

**In scenario 3**: BA evaluation among farms that apply to convert to organic farming, no controls. Instead of controls, national monitoring data (standardized inventories at a landscape scale) or opportunistic citizen science data (should such data exist at these localities) can be used as background time series. Although background data and BA-inventory data may be collected at different spatial scales, this approach can still be useful to contrast changes at organic farms against large-scale population changes of species at the landscape level.

| | Organic | Control | Data | Design | Evidence |
|---|---|---|---|---|---|
| Scenario 1 | Prospective organic farms | Prospective organic farms | BA inventories of organic and control farms | Experimental | Strong causal inference possible |
| Scenario 2 | | Existing conventional farms | BA inventories of organic and conventional farms. | BACI | Moderate causal inference possible |
| Scenario 3 | | Landscape scale monitoring | BA inventories of organic farms. Monitoring data for conventional farmland. | BA with background contrast | Weak causal inference, but allows contrasting BA change against BA trends at the landscape scale |

2010, who found no link between funding source and causal language).

At this point, we want to encourage the multiple actors involved in conservation biology and similar disciplines working with impact evaluations of environmental interventions to improve the scientific rigour with which studies are reported and discussed. While it may seem an intimidating challenge to get authors to openly discuss limitations to their studies, the recognition and discussion of potentially important limitations by authors represent a crucial part of the scientific discourse and will benefit the scientific community and other users of the evidence. Here, a great deal of responsibility lies with the editors of scientific journals to make certain that peer-reviewers also review papers in terms of their internal validity. As an example, research articles in social sciences frequently include a dedicated, and mandatory, limitations section as part of the general discussion. As suggested by Puhan et al. (2012) in the field of biomedicine, we highlight discussing limitations of impact evaluations more transparently, including different sources of bias and the type of

information that would be important to provide for reasons of the scientific method. Further, to increase the legitimacy and quality of systematic reviews, environmental systematic reviews should pay more attention to the internal validity of evidence used, especially relating to unaccounted selection bias. This is something that the research community must do together, in collaboration and with the support of funding agencies, policymakers, and the editors of scientific journals.

## AUTHOR CONTRIBUTIONS

TP suggested the subject area, and TP, JJ, and MH designed the data collection protocol and cross-validated the data collected. JJ compiled and summarized all data. JJ, TP, and MH wrote the first drafts of the manuscript. All authors read their share of original papers to be included in the study, discussed the design of the study, and compiled the data to be included. The subultimate version of the manuscript was written by JJ and TP while all other authors commented on that version, and JJ and TP finalized the manuscript. The revised manuscript was written largely by TP and JJ with important contributions by JKn, DA, and AA.

## ACKNOWLEDGEMENTS

This study was supported by grants from the Swedish Research council (VR) (to ML, JKn) and Swedish research council for sustainable development (to TP, AA, DA, JKn, ML, MP-M, DR).

## DATA AVAILABILITY STATEMENT

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files.

## REFERENCES

Ansell, D., Freundenberger, D., Munro, N., & Gibbons, P. (2016). The cost-effectiveness of agri-environment schemes for biodiversity conservation: A quantitative review. *Agriculture, Ecosystems and Environment*, *225*, 184–191. https://doi.org/10.1016/j.agee.2016.04.008

Balmford, A., Amano, T., Bartlett, H., Chadwick, D., Collins, A., Edwards, D., … Eisner, R. (2019). The environmental costs and benefits of high-yield farming. *Nature Sustainability*, *1*, 477–485. https://doi.org/10.1038/s41893-018-0138-5

Baylis, K., Honey-Rosés, J., Börner, J., Corbera, E., Ezzine-de-Blas, D., Ferraro, P. J., … Wunder, S. (2015). Mainstreaming impact evaluation in nature conservation. *Conservation Letters*, *9*(1), 58–64. https://doi.org/10.1111/conl.12180

Berthet, E. T., Bretagnolle, V., Lavorel, S., Sabatier, R., Tichit, M., & Segrestin, B. (2018). Applying ecological knowledge to the innovative design of sustainable agroecosystems. *Journal of Applied Ecology*, *56*, 44–51. https://doi.org/10.1111/1365-2664.13173

Brooks, T. M., Mittermeier, R. A., da Fonseca, G., Gerlach, J., Hoffmann, M., Lamoreux, J. F., … Rodrigues, A. S. L. (2006). Global biodiversity conservation priorities. *Science*, *313*(5783), 58–61. https://doi.org/10.1126/science.1127609

Bull, J. W., Gordon, A., Law, E. A., Suttle, K. B., & Milner-Gulland, E. J. (2014). Importance of baseline specification in evaluating conservation interventions and achieving no net loss of biodiversity. *Conservation Biology*, *28*(3), 799–809. https://doi.org/10.1111/cobi.12243

Burns, E., Zammit, C., Attwood, S. J., & Lindenmayer, D. B. (2016). The environmental stewardship program. Lessons on creating long-term agri-environment schemes. In D. H. Ansell, F. G. Gibson, & D. J. Salt (Eds.), *Learning from agri- environment schemes in Australia: Investing in biodiversity and other public goods and services in farming landscapes* (pp. 33–51). Canberra, Australia: ANU Press.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., … Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59–67. https://doi.org/10.1038/nature11148

Chevalier, M., Russell, J. C., & Knape, J. (2019). New measures for evaluation of environmental perturbations using before-after-control-impact analyses. *Ecological Applications*, *29*(2), e01838–13. https://doi.org/10.1002/eap.1838

Christie, A. P., Amano, T., Martin, P. A., Shackelford, G. E., Simmons, B. I., & Sutherland, W. J. (2019). Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *Journal of Applied Ecology*, *56*(12), 2742–2754. https://doi.org/10.1111/1365-2664.13499

Cofield, S. S., Corona, R. V., & Allison, D. B. (2010). Use of causal language in observational studies of obesity and nutrition. *Obesity Facts*, *3*(6), 353–356. https://doi.org/10.1159/000322940

Collaboration for Environmental Evidence. (2013). Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. Environmental evidence: Retrieved from www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf

de Loma, T. L., Osenberg, C. W., Shima, J. S., Chancerelle, Y., Davies, N., Brooks, A. J., & Galzin, R. (2008). A framework for assessing impacts of marine protected areas in Moorea (French Polynesia) 1. *Pacific Science*, *62*(3), 431–441. https://doi.org/10.2984/1534-6188(2008)62[431:AFFAIO]2.0.CO;2

Eken, G., Bennun, L., Brooks, T. M., Darwall, W., Fishpool, L., Foster, M., … Tordoff, A. (2004). Key biodiversity areas as site conservation targets. *Bioscience*, *54*(12), 1110–1118. https://doi.org/10.1641/0006-3568(2004)054%5B1110:KBAASC%5D2.0.CO;2

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*(1), 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Eyhorn, F., Muller, A., Reganold, J. P., Frison, E., Herren, H. R., Luttikholt, L., … Smith, P. (2019). Sustainability in global agriculture driven by organic farming. *Nature Sustainability*, *2*, 253–255. https://doi.org/10.1038/s41893-019-0266-6

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

Ferraro, P. J., & Pattanayak, S. K. (2006). Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *PLoS Biology*, *4*(4), e105–7. https://doi.org/10.1371/journal.pbio.0040105

Fisher, B., Balmford, A., Ferraro, P. J., Glew, L., Mascia, M., Naidoo, R., & Ricketts, T. H. (2013). Moving Rio forward and avoiding 10 more years with little evidence for effective conservation policy. *Conservation Biology*, *28*(3), 880–882. https://doi.org/10.1111/cobi.12221

França, F., Louzada, J., Korasaki, V., Griffiths, H., Silveira, J. M., & Barlow, J. (2016). Do space-for-time assessments underestimate the impacts of logging on tropical biodiversity? An Amazonian case study using dung beetles. *Journal of Applied Ecology*, *53*(4), 1098–1105. https://doi.org/10.1111/1365-2664.12657

Gabriel, D., Carver, S. J., Durham, H., Kunin, W. E., Palmer, R. C., … Benton, T. G. (2009). The spatial aggregation of organic farming in England and its underlying correlates. *Journal of Animal Ecology*, *46*(2), 323–333.

Groves, C. R., Jensen, D. B., Valutis, L. L., Redford, K. H., Shaffer, M. L., Scott, J. M., … Anderson, M. G. (2002). Planning for biodiversity conservation: Putting conservation science into practice. *Bioscience*, *52*(6), 499–512. https://doi.org/10.1641/0006-3568(2002)052[0499:PFBCPC]2.0.CO;2

Haddaway, N. R., & Bilotta, G. S. (2016). Systematic reviews: Separating fact from fiction. *Environment International*, *92–93*(C), 578–584. https://doi.org/10.1016/j.envint.2015.07.011

Kehinde, T., & Samways, M. J. (2014). Effects of vineyard management on biotic homogenization of insect-flower interaction networks in the Cape Floristic Region biodiversity hotspot. *Journal of Insect Conservation 18*, 469–477. https://doi.org/10.1007/s10841-014-9659-z

Kleijn, D., & Sutherland, W. (2003). How effective are European agri-environment schemes in conserving and promoting biodiversity? *Journal of Applied Ecology*, *40*(6), 947–969.

Lipton, R., & Ødegaard, T. (2005). Causal thinking and causal language in epidemiology: It's in the details. *Epidemiologic Perspectives & Innovations*, *2*, 8. https://doi.org/10.1186/1742-5573-2-8

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, *21*(1), 121–145. https://doi.org/10.1146/annurev.publhealth.21.1.121

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

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*(6772), 853–858. https://doi.org/10.1038/35002501

Nomura, H., Yabe, M., Nishio, T., Izumi, M., Hirai, K., & Kurokawa, T. (2013). Framework for improvement of farmland biodiversity in Japan, *Journal of Environmental Planning and Management*, *56*(5), 743–758. https://doi.org/10.1080/09640568.2012.702100

Osenberg, C. W., Shima, J. S., Miller, S. L., & Stier, A. C. (2011). ECOLOGY: Assessing effects of marine protected areas: Confounding in space and possible solutions. In J. Claudet (Ed.), *Assessment of the effectiveness of marine protected areas* (Part III, pp. 143–167). Cambridge, UK: Cambridge University Press.

Puhan, M. A., Akl, E. A., Bryant, D., Xie, F., Apolone, G., & ter Riet, G. (2012). Discussing study limitations in reports of biomedical studies-the need for more transparency. *Health and Quality of Life Outcomes*, *10*(1), 23. https://doi.org/10.1186/1477-7525-10-23

Pullin, A. S., & Knight, T. M. (2001). Effectiveness in conservation practice: Pointers from medicine and public health. *Conservation Biology*, *15*(1), 50–54. https://doi.org/10.1046/j.1523-1739.2001.99499.x

Pullin, A. S., & Knight, T. M. (2009). Doing more good than harm – Building an evidence-base for conservation and environmental management. *Biological Conservation*, *142*(5), 931–934. https://doi.org/10.1016/j.biocon.2009.01.010

Reganold, J. (2016). Can we feed 10 billion people on organic farming alone? *The Guardian*. https://www.theguardian.com/sustainable-business/2016/aug/14/organic-farming-agriculture-world-hunger

ter Riet, G., Chesley, P., Gross, A. G., Siebeling, L., Muggensturm, P., Heller, N., … Puhan, M. A. (2013). All that glitters isn't gold: A survey on acknowledgment of limitations in biomedical studies. *PLoS ONE*, *8*(11), e73623–6. https://doi.org/10.1371/journal.pone.0073623

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. (2007). The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal*, *44*(2), 400–413. https://doi.org/10.3102/0002831207302174

Savage, S. (2015). The lower productivity of organic farming: A new analysis and its big implications. *Forbes*. Retrieved from https://www.forbes.com/sites/stevensavage/2015/10/09/the-organic-farming-yield-gap

Science for Environment Policy. (2017). Agri-environmental schemes: how to enhance the agriculture-environment relationship. Thematic Issue 57. Issue produced for the European Commission DG Environment by the Science Communication Unit, UWE, Bristol.

Stubbs, M. (2013). Conservation reserve program (CRP): Status and current issues. Congressional Research Service Report for Congress. No. R42783.

Sutherland, W. J., Pullin, A. S., Dolman, P. M., & Knight, T. M. (2004). The need for evidence-based conservation. *Trends in Ecology & Evolution*, *19*(6), 305–308. https://doi.org/10.1016/j.tree.2004.03.018

Underwood, A. J. (1992). Beyond BACI—the detection of environmental impacts on populations in the real, but variable, world. *Journal of Experimental Marine Biology and Ecology*, *161*(2), 145–178. https://doi.org/10.1016/0022-0981(92)90094-Q

Underwood, A. J. (1997). *Experiments in ecology*. Cambridge, UK: Cambridge University Press.

Wikberg, S., Perhans, K., Kindstrand, C., Djupström, L. B., Boman, M., Mattsson, L., … Gustafsson, L. (2009). Cost-effectiveness of conservation strategies implemented in boreal forests: The area selection process. *Biological Conservation*, *142*(3), 614–624. https://doi.org/10.1016/j.biocon.2008.11.014

## SUPPORTING INFORMATION

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

**How to cite this article:** Josefsson J, Hiron M, Arlt D, et al. Improving scientific rigour in conservation evaluations and a plea deal for transparency on potential biases. *Conservation Letters*. 2020;13:e12726. https://doi.org/10.1111/conl.12726