POLICY FORUM

ENVIRONMENTAL POLICY

Create a culture of experiments in environmental programs

Organizations need a better “learning by doing” approach

By **Paul J. Ferraro[1,2], Todd L. Cherry[3], Jason F. Shogren[3], Christian A. Vossler[4], Timothy N. Cason[5], Hilary Byerly Flint[6], Jacob P. Hochard[6], Olof Johansson-Stenman[7], Peter Martinsson[8], James J. Murphy[9], Stephen C. Newbold[3], Linda Thunström[3], Daan van Soest[10], Klaas van 't Veld[3], Astrid Dannenberg[11], George F. Loewenstein[12], Leaf van Boven[13]**

An understanding of cause and effect is central to the design of effective environmental policies and programs. But environmental scientists and practitioners typically rely on field experience, case studies, and retrospective evaluations of programs that were not designed to generate evidence about cause and effect. Using such methods can lead to ineffective or even counterproductive programs. To help strengthen inferences about cause and effect, environmental organizations could rely more on formal experimentation within their programs, which would leverage the power of science while maintaining a “learning by doing” approach. Although formal experimentation is a cornerstone of science and is increasingly embedded in nonenvironmental social programs, it is virtually absent in environmental programs. We highlight key obstacles to such experimentation and suggest opportunities to overcome them.

By “formal experimentation,” we mean the deliberate creation of spatial or temporal variation in program implementation with the intent of quantifying impacts and elucidating mechanisms. For example, consider an environmental agency that wants to learn how best to encourage polluters to comply with environmental regulations. Instead of implementing a single change in auditing practices across all polluting facilities, the agency could randomly vary implementation of two auditing practices and contrast how facilities respond (see the first figure) [for an analogous real-world example, see (*1*)]. By creating deliberate variation in how programs are implemented, program administrators can more easily learn about the features that make programs effective. Although experimentation in natural resource management has a long history, including in the context of adaptive management, we focus on embedding experiments in the implementation of policies or programs that affect human behavior. For example, in a not-atypical type of environmental policy experiment that tests whether thinning a reforested plot leads to more harvestable timber, human behavior is controlled by the experimentalist, whereas in a much less common type of experiment that tests alternative design features of a program that encourages more reforestation behavior, human behavior is endogenous and uncertain.

Despite the benefits of adding experimental variation to program implementation, as demonstrated in nonenvironmental contexts such as health and education, environmental organizations rarely do so. Consider two US federal agencies with substantial environmental program portfolios: the US Environmental Protection Agency (USEPA) and the US Department of Agriculture (USDA). In the past 30 years, each has embedded formal experimentation in their environmental programs fewer than a half dozen times. In Europe, we know of only a single example of formal experimentation embedded within government-implemented environmental programs (*2*). Formal experimentation is similarly almost nonexistent among nongovernmental and multilateral environmental organizations. Although environmental actors engage in thousands of informal “experiments” every year (such as pilot programs), these are not designed to test the implicit hypotheses that justify the implementation of current programs or understand how to make these programs more effective.

Formal experimentation in environmental programs is absent because science typically stops when implementation starts. Over the past five decades, governmental and nongovernmental actors have invested substantial resources to understand the status and trends of myriad environmental indicators. These investments have been motivated by scientific uncertainty about how complex environmental systems function and by a recognition that reducing this uncertainty is critical to designing effective programs.

Yet uncertainty also plagues program efficacy. The coupled natural-human systems in which environmental programs are implemented are complex, and our understanding of how programs influence the trajectory of these systems is incomplete. When new program designs in nonenvironmental contexts are assessed through formal experimentation, proponents often learn that the innovations fail to have the intended effects (*3*). Scientists and practitioners should not expect innovations in environmental programs to be any different.

The absence of experimentation within environmental programs can be explained in part by historical reasons. Compared with other social policy fields such as health, poverty, and education, the environmental policy field is much younger and would be expected to be a late adopter of innovative ways of generating evidence. Moreover, the human benefits from effective environmental programs are less salient than in other social policy fields. The foregone benefits from ineffective programs are also less salient, putting less pressure on program staff to show effectiveness. Last, environmental practice is dominated by lawyers, engineers, and natural and physical scientists who—unlike health, behavioral, and social scientists—do not typically use experimental designs in real-world contexts and may not anticipate complex human responses to what seem like straightforward policy and program decisions. Yet there are no structural barriers to experimentation in the environmental field.

CONCERNS ABOUT EXPERIMENTATION

Four primary concerns about embedding formal experimentation into environmen-

[1]Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD, USA. [2]Carey Business School, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Economics, University of Wyoming, Laramie, WY, USA. [4]Department of Economics, University of Tennessee, Knoxville, TN, USA. [5]Department of Economics, Purdue University, West Lafayette, IN, USA. [6]Haub School of Environment and Natural Resources, University of Wyoming, Laramie, WY, USA. [7]Department of Economics, University of Gothenburg, Gothenburg, Sweden. [8]Department of Technology, Management and Economics, Technical University of Denmark, Kongens Lyngby, Denmark. [9]Department of Economics, University of Alaska-Anchorage, Anchorage, AK, USA. [10]Department of Economics, Tilburg University, Tilburg, The Netherlands. [11]Institute of Economics, University of Kassel, Kassel, Germany. [12]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA. [13]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO, USA.Email: pferraro@jhu.edu

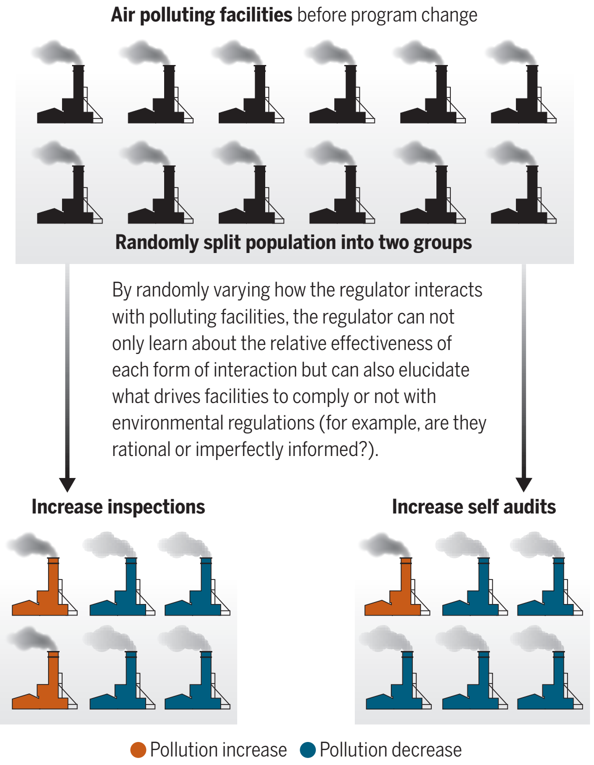
tal programs need to be addressed: delayed action, feedback lags, structural barriers, and ethical questions.

First, identifying ways to create experimental variation and measure outcomes can delay scaled-up implementation, letting environmental damages accumulate. Yet for the case of ineffective programs, the accumulated damages could be much larger when program managers rely on retrospective evaluations that use nonexperimental, postimplementation data. The costs of delays from experimentation will depend on how effectively the program meets its objectives and how quickly damages accumulate. In some cases, large-scale action may be required without waiting for experimentation (akin to "emergency authorizations" in medicine). Yet we believe that in many cases, experimentation embedded in program implementation will improve outcomes in the long run, even at the cost of some delay in the short run. Similar arguments have been made in the recent COVID-19 pandemic, in which calls for quick action and for rigorous evidence seemed to be in opposition (*4*).

Second, the full effects of a program may not materialize for many years (for example, long-run climate impacts), and the evidence may no longer be useful by the time it is available. Yet for many environmental problems, the culprit is human behavior, for which the desired changes can be measured on shorter timescales (for example, changes in energy consumption by households or fertilizer use by farmers). Measures of short-term environmental indicators along the hypothesized causal path may also help elucidate whether the intervention is working as intended (for example, measure pollutants that change relatively rapidly rather than health conditions that change more slowly).

A culture of experimentation within environmental programs

To reduce pollution, regulators can increase on-site inspections, or they can increase opportunities for facilities to do self audits, with some penalty leniency when violations are self reported. Self audits may be less effective at reducing pollution (measured remotely) than on-site inspections because self audits allow facilities to hide their noncompliance. Yet self audits may be more effective because they make facilities more aware about the law and its relationship to their operations and because they transform errors of omission into errors of commission.


Air polluting facilities before program change
Randomly split population into two groups
Increase inspections
Increase self audits
● Pollution increase
● Pollution decrease

By randomly varying how the regulator interacts with polluting facilities, the regulator can not only learn about the relative effectiveness of each form of interaction but can also elucidate what drives facilities to comply or not with environmental regulations (for example, are they rational or imperfectly informed?).

GRAPHIC: N. CARY/*SCIENCE*

Third, structural constraints, such as legal and regulatory rules, may present barriers to experimentation. The degree to which such barriers exist, however, is difficult to ascertain given that there has been so little historical effort allocated to experimentation.

A fourth concern may seem on the surface to be the most problematic: Opponents of experimentation question the ethics of treating some people (or nonhuman organisms or ecological communities) differently than others (*5*). This concern arises from a presumption that those exposed to a program, or a specific version of it, are sure to benefit from it. That assumption, however, is not necessarily true. The effects of many environmental programs are uncertain.

One could argue that environmental organizations have an ethical obligation to better understand the effects of untested programs, or changes in programs, before large groups of humans and other species, particularly vulnerable subgroups, are exposed to them (akin to the principle of "equipoise," a state of genuine uncertainty about the comparative merits of different approaches, which is the ethical basis for justifying randomized treatments in medical trials). Even programs that do not directly harm the environment or people may simply be ineffective. Directing resources to ineffective interventions has substantial ethical implications, especially for environmental problems that are time sensitive, such as the loss of biological diversity and the accumulation of persistent pollutants.

If environmental organizations were guided by an ethical precept that required evidence before changing or scaling up a program, the science and practice of environmental protection would look different and be more successful. Environmental programs would routinely be subjected to experimentation that deliberately manipulates the temporal and spatial variability of implementation. Program managers, perhaps in collaboration with academics, would then evaluate the results to better understand the consequences, intended and unintended, of the variations in implementation. This evidence would provide opportunities to adjust and improve current and future programs (*6*, *7*). This cycle of program innovation, experimentation, learning, and adaptation is a hallmark of evidence-based programs in other fields.

ENCOURAGING EXPERIMENTATION

Although the constraints on engaging in experimentation will vary by organization, the opportunities for experimentation have some commonalities. On the basis of experiences in other social policy fields, we offer four recommendations for expanding the opportunities for experimentation in environmental programs [for others, see (*8*, *9*)].

Political and legal simplicity

Running an experiment that contrasts an entire program to a no-program control may require extensive legal and political approvals, as well as expose implementers to reputational risks and coordination costs. Instead, one version of program implementation can be compared with another version by manipulating program attributes for which managers already have the authority to change (often called "A/B testing" in the private sector). For example, program managers could contrast the effects on pollution compliance from on-site inspections (status quo) versus remote inspections. Leveraging already-planned pilot programs can also be a practical way to facilitate learning when the pilot's implementation is varied across space or time in ways unrelated to the program's target outcomes.

Financial simplicity

Given that the additional costs of experimentation largely come from the costs of measuring outcomes, organizations can focus on contexts in which the outcomes are collected as part of program operations (such as pollution discharges) or are publicly available (such as satellite data of land use).

Learning focused

To achieve higher returns on investment, organizations should focus on experimentation that yields results that can be generalized across multiple programs. Generalizability is more plausible when the program features being manipulated are found in many programs (such as capacity building and incentives) or motivated by similar theories of change.

Partnership enhanced

A quick, inexpensive way for environmental organizations to acquire the technical capacity to design and analyze experiments, while keeping the operations in-house, is to embed trained experimentalists from outside the organization (for example, through federal Voluntary Service Agreements in the US context).

Strengthening the culture of experimentation in the environmental community will require changes in norms and incentives. Program managers are often not rewarded for evidence about program effectiveness but rather for achieving other objectives (such as moving money to constituents, avoiding litigation by private actors, or pleasing funders). Nevertheless, changes in norms and incentives are occurring. One recent example of change is the creation of "behavioral insights teams" in governmental and multilateral organizations. These teams help program managers to formally experiment with program changes inspired by insights from the behavioral sciences (*10*).

For federal agencies in the United States, changes in norms and incentives are also occurring through the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act). The Evidence Act and complementary memoranda from the Executive branch encourage a culture of experimentation both directly and indirectly. They encourage experimentation directly by emphasizing the power and political acceptability of randomized implementation designs (*11–14*). They encourage experimentation indirectly by requiring agencies to create annual learning agendas and a strategy and budget to meet their agenda objectives. Learning agendas comprise a set of questions that, when answered, are expected to have the biggest impact on an agency's performance. Yet the Evidence Act and its associated guidance do not provide explicit rewards to staff for posing substantive learning questions and using experimentation to generate high-quality answers to these questions. Thus, by itself, the Evidence Act may be insufficient to create a meaningful culture of experimentation within environmental agencies.

One way to further foster a culture of experimentation and embed learning in daily operations among US federal agencies would be through a new executive order (EO) similar in spirit to EO 12291 for Cost-Benefit Analyses. This new EO would be triggered if a new environmental program, or change in a current program, were to exceed a size threshold, which could be measured by program funding or the size of the affected population. The EO would require the implementing agency to first ascertain the equipoise of the proposed program or change in program: Is there strong empirical evidence that the proposed action is the best option? If not, then the agency would be required to embed experimentation into the program with the intent of quantifying environmental and social impacts and understanding the mechanisms through which those impacts arise. The EO would require that agencies insert a step between proposing a programmatic change and scaling that programmatic change up to the entire eligible population. The EO would also encourage environmental agency staff to involve statisticians and behavioral scientists before implementation. Currently, if these experts are called on at all, it is after implementation to assess what may have transpired—a challenging task when implementation was not designed to generate evidence about impacts and mechanisms. In addition to characterizing what type of experimentation is acceptable, the EO would also have a stopping rule, similar in spirit to stopping rules used to decide when to end medical treatment trials. Likewise, the EO would also define when it may be acceptable to forego experimentation.

Four conditions when experimentation pays off

Pre-Change Ambiguity

When theory and experience alone cannot unambiguously predict the expected impacts ofchanges in program implementation

Post-change Ambiguity

When estimating counterfactual outcomes in the absence of a change in program implementation is challenging using traditional approaches

High Implementation Cost

When a change in program implementation is unlikely to pass a benefit-cost test, or cost-effectiveness assessment, without medium or large impacts

Generalizability of Results

When the lessons learned from experimentation are generalizable beyond the context in which the program change was implemented.

Scientists and practitioners can legitimately argue about the benefits and opportunity costs of allocating scarce time and financial resources to formal experimentation in the environmental sector. Should half of environmental programs include experimentation? Is 10% the right amount? Although the optimal share is debatable, we believe that the current allocation of roughly 0% is suboptimal. How much experimentation is embedded in programs should depend on contextual attributes that make experimentation most valuable (see the second figure).

We recognize that experimentation is not the only way that a scientific lens can be applied to improve our understanding of program implementation. Experimentation is best viewed as part of a mixed-methods approach to generating evidence rather than as a substitute for more traditional ways of gathering evidence. Experimentation should, however, be a regular feature of programs, not a rarity. ■

REFERENCES AND NOTES

1. D. I. Levine, M. W. Toffel, M. S. Johnson, *Science* **336**, 907 (2012).
2. K. Telle, *J. Public Econ*. **99**, 24 (2013).
3. A. Ventures, *Straight Talk on Evidence*, 13 April 2018); https://www.straighttalkonevidence.org/2018/04/13/how-to-solve-u-s-social-problems-when-most-rigorous-program-evaluations-find-disappointing-effects-part-two-a-proposed-solution.
4. A. J. London, J. Kimmelman, *Science* **368**, 476 (2020).
5. E. L. Pynegar, J. M. Gibbons, N. M. Asquith, J. P. G. Jones, *Oryx* **55**, 235 (2019).
6. Moving to Opportunity (MTO) for Fair Housing Demonstration Program, https://www2.nber.org/mtopublic/.
7. R. H. Brook *et al*., "The Health Insurance Experiment: A classic RAND study speaks to the current health care reform debate" (RAND Corporation, 2006).
8. J. Fox, *Evidence Matters*, 29 May 2019; https://www.3ieimpact.org/blogs/how-can-rethink-lessons-field-experiments-inform-future-research-transparency-participation.
9. E. Duflo, *Am. Econ. Rev*. **107**, 1 (2017).
10. S. Wendel, *Behavioral Scientist*, 5 October 2020); https://behavioralscientist.org/who-is-doing-applied-behavioral-science-results-from-a-global-survey-of-behavioral-teams.
11. https://www.whitehouse.gov/omb/information-for-agencies/evidence-and-evaluation/.
12. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/memorandum-on-restoring-trust-in-government-through-scientific-integrity-and-evidence-based-policymaking.
13. https://www.whitehouse.gov/wp-content/uploads/2021/06/M-21-27.pdf.
14. Appendix A of OMB M-19-23 describes four broad types of evidence that agencies should use as they implement the Evidence Act: foundational fact finding, policy analysis, program evaluation, and performance measurement. This guidance goes a step further to specify the broad range of methodological approaches that agencies should consider. These approaches include, but are not limited to, "pilot projects, randomized controlled trials, quantitative survey research and statistical analysis, qualitative research, ethnography, research based on data linkages in which records from two or more datasets that refer to the same entity are joined, well-established processes for community engagement and inclusion in research, and other approaches that may be informed by the social and behavioral sciences and data science.

10.1126/science.adf7774