Ferraro, P. J. (2009). Counterfactual thinking and impact evaluation in environmental policy. In M. Birnbaum & P. Mickwitz (Eds.), *Environmental program and policy evaluation. New Directions for Evaluation, 122,* 75–84.

7

Counterfactual Thinking and Impact Evaluation in Environmental Policy

Paul J. Ferraro

Abstract

Impact evaluations assess the degree to which changes in outcomes can be attributed to an intervention rather than to other factors. Such attribution requires knowing what outcomes would have looked like in the absence of the intervention. This counterfactual world can be inferred only indirectly through evaluation designs that control for confounding factors. Some have argued that environmental policy is different from other social policy fields, and thus attempting to establish causality through identification of counterfactual outcomes is quixotic. This chapter argues that elucidating causal relationships through counterfactual thinking and experimental or quasi-experimental designs is absolutely critical in environmental policy, and that many opportunities for doing so exist. Without more widespread application of such approaches, little progress will be made on building the evidence base in environmental policy. © Wiley Periodicals, Inc.

In presentations on evaluation to environmental scientists and field practitioners, I often begin by describing a Florida conservation education program that was designed to reduce residential outdoor water use (Mulville-Friel & Anderson, 1996). The program was piloted in a neighborhood where outdoor water consumption was measured before and after

Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/ev.297

the program. After program implementation, mean household water consumption declined 29%. The effort was expanded to another neighborhood where water consumption declined 38%.

Most audience members indicate that a program officer would be justified in broadly scaling up the education program. In their eyes, the evaluation strategy has the required attributes to credibly evaluate program effectiveness: a clear theory of change and a specific and measurable outcome indicator observed before and after the intervention. They also indicate that the quality of the program's evidence is much better than in many environmental programs. Environmental programs often lack clear theories of causal relationships and baseline indicators. I end the Florida example by asking audience members what they would think about the evidence of program effectiveness if I told them (truthfully) that rainfall had increased at the time when the program was expanded to a second community, and that water consumption declined 31% in a nearby community that received no education. They immediately realize that the original evaluation design was missing important elements.

Impact evaluations assess the degree to which changes in outcomes can be attributed to a program, policy, or intervention "treatment," rather than to confounding factors that also affect the outcomes. Impact evaluations answer the question, "Does the intervention work better than no intervention at all (or a proposed alternative intervention)?" An answer requires knowing what outcomes would have looked like in the absence of the intervention. This counterfactual world, however, can be inferred only indirectly. Efforts to measure causes of environmental outcomes through counterfactual thinking are rare in the environmental literature.

In this chapter, I argue that counterfactual thinking is critical to building the evidence base in environmental policy about what types of interventions work and under what conditions. More specifically, I argue that:

- Much of what is called evaluation of environmental program impact is simply monitoring of indicators
- Counterfactual thinking is essential to drawing inferences about program effectiveness because realistic behavioral theories typically yield ambiguous predictions about environmental program impacts, and because environmental outcomes are affected by many confounding factors correlated with the timing and location of interventions
- Advancing counterfactual thinking through experimental and quasi-experimental designs faces the same barriers that exist in other social policy fields, but these barriers are particularly pervasive in environmental policy
- In comparison to other social policy fields, there are few examples of experimental or quasi-experimental evaluations
- Despite the barriers and paucity of examples, there are substantial opportunities to elucidate causal relationships through experimental and quasi-experimental designs

The Need for Counterfactual Thinking in Environmental Policy Evaluation

The environmental literature is replete with data on biophysical processes and status indicators. Yet when evaluating how environmental policies and programs affect these processes and indicators, the literature lags behind other social policy fields, such as public health and poverty reduction. For example, a global ecosystem assessment (Millennium Ecosystem Assessment) contained hundreds of pages of data to support its characterization of the state of the world's ecosystems. However, one of its "main messages" (MA, 2005, p. 122) is that "few well-designed empirical analyses assess even the most common biodiversity conservation measures." Others have made similar arguments in the case of ecosystem conservation (Pullin & Knight, 2001; Sutherland, Pullin, Dolman, & Knight, 2004; Saterson et al., 2004; Stem, Margoluis, Salfasky, & Brown, 2005; Ferraro & Pattanayak, 2006), energy conservation (Frondel & Schmidt, 2005), and pollution policy (Bennear & Coglianese, 2005; Greenstone & Gayer, 2007).

Much of the emphasis to date has been on monitoring: collecting data on indicators of status and trend, such as pollution levels, habitat areas, and management effectiveness scores. Indicators are important because they allow us to document progress toward policy goals. They can reveal if more action is needed to achieve a goal. Alone, however, they cannot reveal if an intervention affected progress toward that goal.

Environmental scientists and practitioners often assume that evaluation is simply the act of taking a careful look at the monitoring data. If the indicator improves, a program is deemed to be "working." If the indicator worsens, one infers the program is "failing." In contrast, impact evaluation contrasts changes in an indicator to some estimate of the counterfactual change in the indicator, the change that would have occurred without program. The essence of counterfactual thinking is elimination of plausible rival interpretations of observed outcomes.

Hidden Biases in Environmental Program Evaluation

Eliminating plausible rival interpretations of observed outcomes requires complex theories. Theories have power in the degree to which they exclude what one can observe and still find the theory to be correct. Demonstrating that no plausible alternative theories can account for what has been observed is important in counterfactual analyses. Theory alone, however, is not sufficient to identify impacts.

Realistic theories of behavioral change in environmental contexts are often consistent with positive, negative, and neutral program impacts. For example, interventions to induce firms and citizens to adopt energy-efficient technologies in order to lower energy consumption and thereby reduce greenhouse gas emissions (IPCC, 2007) may achieve the opposite, particularly in low-income nations (see special issues of *Energy Policy* and *Energy and Environment,* 2000). Although efficient technologies reduce the energy consumption per unit of power, heating, cooling, or lighting, they also reduce the effective price of these outputs and thus increase demand (thermostat settings change, lights stay on longer). Complex potential responses are common in environmental programs that use regulations, incentives, or education (information). Thus the extent to which many environmental programs have the desired effect is an empirical question.

Empirical analyses, however, are made difficult by pervasive confounding factors that mask program failure or mimic program success. This includes (1) cotemporaneous factors that are correlated with the treatment intervention and outcomes; and (2) selection bias, where treated units are selected, or select themselves, to receive the intervention on the basis of characteristics that also affect the outcome. These sources of confounding factors are found in nearly all environmental programs, and predicting their direction and magnitude ex ante is difficult. This in turn confounds efforts at credible ex post impact evaluations.

With regard to cotemporaneous confounding factors, a large set of factors, including changes in weather and in relative prices and other economic characteristics (such as fuel prices or employment opportunities), affect environmental outcomes. Comparing outcomes in the treatment group to outcomes in a control group can reduce bias from cotemporaneous confounders, but pervasive selection bias implies that the outcome of the average untreated observation will rarely represent the counterfactual outcome of the average treated observation. For example, the characteristics that lead program administrators to target certain individuals, firms, species, or areas are frequently correlated with outcomes (see Figure 7.1). Voluntary programs also suffer from self-selection bias. For example, incentive programs (payments for environmental services, eco-labeling, adoption of environmental management systems) often reward people or firms for not engaging in environmentally destructive activities that, at many places and times, would not be done even in the absence of the program.

To distinguish between program effects and hidden biases, counterfactual thinking is absolutely essential. To engage in such thinking, analysts often collect data within quasi-experimental and experimental designs. These designs attempt to identify exogenous variation in the program in order to identify its impact. Greenstone and Gayer (2007) argue that the field of environmental policy is “flush with opportunities to apply these techniques.” In the next section, I demonstrate that there are indeed opportunities for applying these designs.

Nevertheless, such designs are much rarer in environmental policy than in other social policy fields. Their rarity arises from environmental practitioners and scientists’ lack of familiarity with appropriate designs and methods, as well as barriers to doing *any* environmental program impact evaluation, whether through experimental or nonexperimental designs.

Figure 7.1. Experimental and Quasi-Experimental Designs in Environmental Program Evaluation

How much do indoor air pollution reductions affect infant and mother's health? Households that invest to improve indoor air quality also tend to be wealthier, be better informed, and have greater concerns about their health. To address potential confounders, the RESPIRE study (Diaz et al., 2007) randomly provided less-polluting cook stoves to women with children. Women and children in the treatment group experienced substantial reductions in carbon monoxide exposure and reported poor health compared to the control group.

By how much do protected area systems reduce deforestation? Global efforts to reduce tropical deforestation rely on protected areas, which are sited on the basis of characteristics correlated with deforestation. Using spatial data and matching methods to control for observable sources of bias (and potential spillovers from protected to unprotected lands), Andam, Ferraro, Pfaff, Sanchez-Azofeifa, and Robalino (2008) estimate that less than 10% of protected forest in Costa Rica would have been deforested by 1997 in the absence of protection. Conventional before-after-inside-outside estimates, which fail to control for observable sources of bias, overestimate avoided deforestation by 65% or more.

Do protected areas improve local health and incomes? Most studies are based on ex ante predictions from historical use data and strong assumptions, or ex post analyses that prove only that the poor live near protected areas. In contrast, Wilkie et al. (2006) are tracking health and livelihood outcomes of 1,000 households that traditionally used resources around four new Gabonese national parks and 1,000 households that live outside the influence of the same parks.

How does the U.S. Endangered Species Act affect species recovery? Most studies lack a clear counterfactual. Ferraro, McIntosh, and Ospina (2007) use matching methods to select control groups of species and estimate how species listed and funded under the act would have fared had they not been listed or funded. The analysis suggests that the act improves outcomes for species only when accompanied by substantial species-specific funding but makes outcomes worse when a species is listed under the act with little or no funding.

How does mandatory reporting affect firms' propensity to violate standards? Information disclosure regulations are increasingly common, but their effects on the behavior of regulated firms are unclear. Bennear and Olmstead (2008) evaluate the impact of a 1996 amendment to the U.S. Safe Drinking Water Act that requires community drinking water suppliers to mail information about water quality to their customers. Using a difference-in-differences panel data design and a regression discontinuity design, they estimate that the amendment reduced the number of health and other violations substantially, by about 50%.

These barriers include nonlinear response outcomes, such as thresholds; high natural rate of outcome variability; treatments that comprise multiple interventions; infrequent data sampling, nonexistent baselines, and large measurement error; long time lag between intervention and response; programs with multiple interventions; complex spillover effects, as when deforestation pressures or animals migrate; large spatial scales of ecological processes and environmental interventions, such as landscapes and airsheds; unique treatment units without comparators, such as restricted habitats of endemic species; and small operations budgets (see Hockings et al. in this issue).

Although such barriers are found in all social policy fields, they are particularly pervasive in the field of environmental policy. Thus not only do environmental practitioners and scientists have little familiarity with appropriate designs but their field is one of the most difficult evaluation settings in which to apply such designs. However, many environmental programs are aimed at affecting human behavior in the short run, either as individuals or as collectives in the form of communities, governments, and firms. From this perspective, environmental programs are not radically different from programs in other social policy fields. Thus, although measuring program impact on environmental outcomes may often be difficult because of the barriers noted above, measuring program impacts on the intermediate impact of behavioral changes can be easier. For example, if hunting is threatening a species, and a program designed to reduce hunting is observed to have no effect on hunting, one could reasonably conclude that the program has had no effect on long-term population dynamics of the threatened species (the converse, however, would not necessarily be true).

Experimental and Quasi-Experimental Designs in Environmental Program Evaluation

Experimental and quasi-experimental designs focus on selecting the best representation of the counterfactual. They allow evaluators to collect data in a way that produces sufficient variation in key variables to allow identification and measurement of program impact on relevant outcome indicators.

Experimental Designs. Experiments induce variation by controlling how the data are collected. The most popular way of inducing this variation is through randomization of the program assignment. Probability thus enters the experiment only through random assignment, a process controlled by the experimenter. Randomization is the most popular way of inducing exogenous variation in an experiment because it is easily understood by analysts and practitioners alike. For the same reasons, I focus on randomization, but there are other ways to induce exogenous variation so that program assignment is not correlated with outcomes (Heckman, 2005).

Randomized experiments are most easily and cheaply implemented when the program has one stage and few treatment variations. They are often possible in the context of pilot programs that leave some people, communities, or sites as controls. For example, in programs with more eligible participants than the budget can support, one can randomly choose participants among the relevant population. For programs that cannot randomly restrict access, a random encouragement design might be possible where members of the target population are encouraged at random to participate. In programs that are phased in over time, one can randomly select which participants receive the treatment first, thereby allowing the later participants to serve as controls for the early participants.

Despite the widespread presence of these facilitating conditions in the field of environmental policy, randomized experiments are rare. Published studies are most prevalent among social psychologists that test the effect of conservation messages on individual behaviors, such as littering and energy consumption (see references in Oskamp & Schultz, 2006). Other experiments situate themselves on the boundary of health and the environment, such as the impact of improved wood stoves on indoor air pollution and household health (see Figure 7.1).

In recent years, the number of current and proposed randomized environmental evaluations has increased. These include experiments to test the impact of (1) conservation education messages on individual and collective behavior in the United States (water consumption, voting behavior); (2) payments for environmental services on forest cover and household welfare in Vietnam and Uganda (using randomization at village level, rather than individual level); (3) adoption of improved wood stoves on health and economic welfare in India, of compact fluorescent light bulb adoption on energy use in Africa (randomly assign subsidies to purchase lights); and (4) third-party shade-grown coffee certification on environmental and economic outcomes in Latin America.

I do not wish to glorify experimental designs, particularly those that use randomization, whose limitations have been widely debated in the literature (for example, ineffective randomization; randomization biases in which the experimental design itself affects behavior; the inability of unaided randomization to estimate the fraction of a population that benefits from a program). Nevertheless, experimental designs can be particularly helpful in contexts in which there are many plausible biases. Moreover, they can serve as a complement to the quasi-experimental designs discussed in the next section and to nonexperimental designs.

Quasi-Experimental Designs. When true experimental designs are not feasible for political, financial, legal, practical, or ethical reasons, a quasi-experimental design might be possible using available data. If executed with appropriate statistical methods and in full recognition of their limitations, such designs can provide better information about causal impact than nonexperimental designs, especially when quantitative data are available. As in experimental designs, reliability depends on the analyst's ability to specify the counterfactual.

Quasi-experimental designs fall into two categories: (1) designs, such as matching and cross-sectional regressions, which assume treatment assignments are affected only by observable variables for which one can collect data and control in the analysis; and (2) designs that assume treatment assignment takes place on variables that are both observable and unobservable to the analyst. The latter include panel data designs (fixed effects models that control for time-invariant unobservables), as well as "natural experiments," which take advantage of situations in which treatment status is determined by nature, politics, an accident, or some other action beyond the researcher's control (including regression discontinuity; Trochim, 1984; and instrumental variable designs). In some cases, natural experiments only mimic a program but allow one to draw inferences about a hypothetical program's impacts. For example, a recession that lowers pollution emissions may allow one to test the health impacts of a policy that explicitly restricts emissions (Chay & Greenstone, 2003), or widespread forest fires may allow one to estimate health impacts of indoor air pollution programs (Jayachandran, 2006). Quasi-experimental designs are often supplemented by testing the sensitivity of results to potential unobservable confounders or by using tests of known effects to detect the presence of hidden bias (Rosenbaum, 2002).

As in all impact evaluations, the validity of the inference rests on the assumption that assignment to treatment and control groups is not related to other determinants of the outcomes. In category (1) above, treatment is assumed to be unrelated to outcomes, conditional on observable characteristics. In category (2), the independence of treatment and outcomes is asserted through an argument based on knowledge of the program implementation or the known relationship between variables (for example, a variable known to affect who is exposed to a program but uncorrelated with the outcome).

In contrast to the paucity of evaluations using experimental designs, there are dozens of evaluations in the environmental field using quasi-experimental designs (see Figure 7.1 for a few examples). Despite the greater number of examples in the literature, quasi-experimental analyses still make up a small proportion of the environmental evaluation literature.

As with all evaluation designs, one must consider not only their internal validity (i.e., whether one is actually estimating a causal relationship rather than hidden biases) but also their construct validity (whether one is actually measuring the outcome and treatment one reports to be measuring) and external validity (whether the results would be the same for other people, places, or times). These issues, however, are largely context-specific rather than design-specific.

The Future of Environmental Program Impact Evaluations

Counterfactual thinking is important in *any* evaluation seeking to identify program impacts. The best way to promote such thinking is through experimental or quasi-experimental designs that attempt to collect data so that an actual treatment effect would be visibly different from the most plausible hidden biases (Rosenbaum, 2002). Most environmental programs will not be amenable to evaluation as true experiments; nor will they be amenable to sophisticated quasi-experimental designs that require advanced statistical methods, large sample sizes, and rich data sets. Indeed, *most* environmental programs cannot be evaluated with such designs.

Nevertheless, even case study analyses are much more informative when exposed to counterfactual thinking and quasi-experimental designs that collect data so that a treatment effect can be distinguished from hidden biases (Yin, 2003). Most environmental programs should, at a minimum, formulate complex theories of change (causal hypotheses with explicit assumptions); make observations, including at the baseline, on a few important indicators of key theoretical assumptions and outcomes (for example, behavioral changes); consider the likelihood that confounding factors are also affecting outcomes (in other words, carefully consider rival explanations of the observed outcomes); and then make informed judgments about how the program can be changed on the basis of the program's own evidence and evidence from analyses with better internal validity done elsewhere.

Not all environmental programs are amenable to experimental or quasi-experimental designs, but surely *some* of the thousands of environmental programs initiated globally every year are. Although challenges to using these designs exist, their use is no more expensive or complicated than the biological and chemical assessments that are routinely used to develop indicators and to improve understanding of environmental processes. The promise of such designs lies in their ability to complement nonexperimental evaluations and intuition. However, until environmental scientists and practitioners become more aware of these designs and have the incentives and capacity to use them, little progress will be made in building the evidence base for environmental policy.

References

Andam, K., Ferraro, P. J., Pfaff, A., Sanchez-Azofeifa, A., & Robalino, J. (2008). Measuring the effectiveness of protected area networks in reducing deforestation. *Proceedings of the National Academy of Sciences, 105*(41).

Bennear, L. S., & Coglianese, C. (2005). Measuring progress: Program evaluation of environmental policies. *Environment, 47*(2), 22–39.

Bennear, L. S., & Olmstead, S.M. 2008. The impacts of the 'Right to Know': Information disclosure and the violation of drinking water standards. *Journal of Environmental Economics and Management, 56*(2), 117–130.

Chay, K. Y., & Greenstone, M. (2003). The impact of air pollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics, 188*(3), 1121–1167.

Diaz, E., Smith-Sivertsen, T., Pope, D. Lie, R. T., Diaz, A., McCracken, J., et al. (2007). Eye discomfort, headache and back pain among Mayan Guatemalan women taking part in a randomized stove intervention trial. *Journal of Epidemiology and Community Health, 61*(1), 74–79.

Ferraro, P. J., McIntosh, C., & Ospina, M. (2007). The effectiveness of the U.S. Endangered Species Act: Econometric analysis using matching methods. *Journal of Environmental Economics and Management, 54*(3), 245–261.

Ferraro, P. J., & Pattanayak, S. K. (2006). Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *PLoS Biology, 4*(4), 482–488.

Frondel, M., & Schmidt, C. M. (2005). Evaluating environmental programs: The perspective of modern evaluation research. *Ecological Economics, 55*(4), 515–526.

Greenstone, M., & Gayer, T. (2007). *Quasi-experimental and experimental approaches to environmental economics.* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1001330.

Heckman, J. J. (2005). Rejoinder: Response to Sobel. *Sociological Methodology, 35*(1), 135–162.

IPCC. (2007). *Climate change 2007: Mitigation of climate change. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge: Cambridge University Press.

Jayachandran, S. (2008). *Air quality and early life mortality: Evidence from Indonesia's wildfires.* http://www.stanford.edu/~jayachan/indo_fires.pdf.

Millennium Ecosystem Assessment (MA). (2005). *Ecosystems and human well-being: Policy responses.* Washington, DC: Island Press.

Mulville-Friel, D., & Anderson, D. (1996, August). Measuring effectiveness. *Florida Water Resources Journal,* 18–20.

Oskamp, S., & Schultz, P. W. (2006). Using psychological science to achieve ecological sustainability. In S. I. Donaldson, D. E. Berger, & K. Pezdek (Eds.), *Applied psychology: New frontiers and rewarding careers* (pp. 81–106). New York: Routledge Press.

Pullin, A. S., & Knight, T. M. (2001). Effectiveness in conservation practice: Pointers from medicine and public health. *Conservation Biology, 15*(1), 50–54.

Rosenbaum, P. (2002). *Observational studies.* New York: Springer-Verlag.

Saterson, K. A., Christensen, N. L., Jackson, R. B., Kramer, R. A., Pimm, S. L., Smith, M. D., & Wiener, J. B. (2004). Effectiveness in conservation practice: Pointers from medicine and public health. *Conservation Biology, 18*, 597–599.

Stem, C., Margoluis, R., Salfasky, N., & Brown, M. (2005). Monitoring and evaluation in conservation: A review of trends and approaches. *Conservation Biology, 19*(2), 295–309.

Sutherland, W. J., Pullin, A. S., Dolman, P. M., & Knight, T. M. (2004). The need for evidence-based conservation. *Trends in Ecology and Evolution, 19*(6), 305–308.

Trochim, W. (1984). *Research design for program evaluation: The regression-discontinuity approach.* Beverly Hills, CA: Sage.

Wilkie, D., Morelli, G., Demmer, J., Starkey, M., Telfer, P., & Steil, M. (2006). Parks and people: Assessing the human welfare effects of establishing protected areas for biodiversity conservation. *Conservation Biology, 20*(1), 247–249.

Yin, R. K. (2003). *Case study research: design and methods.* Thousand Oaks, CA: Sage.

PAUL J. FERRARO is an associate professor of economics in the Andrew Young School of Policy Studies at Georgia State University.