Advances in Measuring the Environmental and Social Impacts of Environmental Programs

Paul J. Ferraro[1] and Merlin M. Hanauer[2]

[1]Department of Economics, Andrew Young School of Policy Studies, Georgia State University, Atlanta, Georgia 30302; email: pferraro@gsu.edu

[2]Department of Economics, Sonoma State University, Rohnert Park, California 94928; email: hanauer@sonoma.edu

Annu. Rev. Environ. Resour. 2014. 39:10.1–10.23

The *Annual Review of Environment and Resources* is online at environ.annualreviews.org

This article's doi:
10.1146/annurev-environ-101813-013230

Keywords

causal, bias, impact evaluation, selection, heterogeneity, mechanisms

Abstract

Inspired by the success of evidence-based medicine, environmental scholars and practitioners have grown enthusiastic about applying a similar evidence-based approach to solve some of the world's most pressing environmental problems. An important component of the evidence-based movement is the empirical evaluation of program and policy impacts. Impact evaluations draw heavily from recent advances in the empirical study of causal relationships—the effect of one thing on another. This review highlights the key components of these advances and characterizes the way in which they contribute to better evaluations of the environmental and social impacts of environmental programs. The review emphasizes that a solid understanding of these advances is required before environmental scholars and practitioners can begin to collect the relevant data, analyze them within credible research designs, and generate reliable evidence about the effectiveness of the myriad proposed solutions to the world's environmental and social problems.

Contents

INTRODUCTION

In the past three decades, scholars have made substantial advances in the empirical study of causal relationships—the effect of one variable on another (1–5). These advances can contribute to better empirical evaluations of the environmental and social impacts of environmental programs, yet they have passed largely unnoticed by environmental scholars and practitioners (6).

In this review, we highlight the key components of these advances: (*a*) creative and transparent ways to identify and eliminate rival explanations for observed empirical patterns; (*b*) precise, transparent definitions of causal effects in terms of potential outcomes, both observable outcomes and unobservable, counterfactual outcomes; (*c*) a focus on design over methods, with a concomitant emphasis on understanding selection (why are some areas, households, villages, or species exposed to a program and others are not?); and (*d*) an obligation to make transparent the assumptions required for causal inference and then to interrogate these assumptions with theory and data. Transparency about assumptions is critical because causal inferences come from a combination of data and untestable assumptions, never from the data alone.

Causal Effects and Potential Outcomes

Scholars describe causal effects through the language of potential outcomes (often called the Rubin or Neyman-Rubin Causal Model). To describe an effect of an antipollution program, D, on a pollutant, Y, in location i, one can think about two potential pollution outcomes: the outcome that would be observed in the presence of the program, $Y_i(1)$, and the outcome that would be observed in the absence of the program, $Y_i(0)$ (i.e., with no program or with an alternative program).[1]

The causal effect of program D at location i at a particular point in time can be defined as the difference in the two potential outcomes (we suppress the time subscript):

$$\delta_i = Y_i(1) - Y_i(0). \qquad 1.$$

In other words, the impact of D at location i is the difference between pollution with D and pollution without D (whatever else in the environment that is unaffected by D evolves the same in both potential states of nature). This way of viewing causal effects highlights two important insights: (*a*) a causal effect of a program is only defined with respect to a well-defined alternative—one must clearly define what $D = 0$ means, and (*b*) causal effects are unit-specific and may be heterogeneous—different locations may respond differently to program D. As we describe below, one's beliefs about this heterogeneity affects the assumptions required to draw inferences and how one interprets the results of an empirical study.

The challenge for scientists is that, at any given moment, we can observe either $Y_i(1)$ or $Y_i(0)$, but not both. The unobservable potential outcome is termed the counterfactual outcome. Although we might not be able to draw inferences about the causal effect of exposure to D for a specific location, we may be able to draw inferences about the distribution of the effect over a relevant population of locations. For example, using the dominant jargon in the causality literature, we can define the average treatment effect, or ATE (treatment = program/cause):

$$\mathrm{ATE} = \mathrm{E}[\mathrm{Y_i(1) - Y_i(0)}] = \mathrm{E}[\mathrm{Y_i(1)}] - \mathrm{E}[\mathrm{Y_i(0)}], \qquad 2.$$

where $E[.]$ denotes the expectation operator from probability theory. In words, the ATE is the expected causal effect of D on Y for a randomly chosen unit i from the relevant population (units could be locations, firms, households, individuals, species, etc.).

Estimating ATEs is easier than estimating individual-level treatment effects, but the fundamental problem of unobservable counterfactual outcomes remains. For treated units exposed to D, we observe only $E[Y_i(1)]$; $E[Y_i(0)]$ is a counterfactual outcome. For untreated units not exposed to D, we observe only $E[Y_i(0)]$; $E[Y_i(1)]$ is a counterfactual outcome.

Impact evaluation focuses its efforts on finding credible ways to estimate these counterfactual outcomes, using what we can observe as a surrogate for what we cannot. In other words, we seek a group of untreated units for which $E[Y_i(0)]$ can represent the counterfactual outcomes of the treated units and a group of treated units for which $E[Y_i(1)]$ can represent the counterfactual outcomes of the untreated units. To estimate the ATE, we seek units for which $E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1]$, where "$\mid D_i = 1$" denotes "conditional on the unit being exposed to the treatment" and "$\mid D_i = 0$" denotes "conditional on the unit being exposed to the untreated, or control, condition" (in words, we seek treated and untreated units for which the mean observed outcome under the control condition for the untreated units is equal to the mean counterfactual outcome for the treated units, had they instead been exposed to the control condition). We also seek units for which $E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0]$.

In most cases, however, we do not want to know the effect of a cause on a randomly chosen unit from the population. We wish to know the causal effect on a randomly chosen unit that was

[1] We assume that D is binary—it is either present ($D = 1$) or not present ($D = 0$). Allowing the program to vary in composition or strength simply generates more potential outcomes (e.g., pollutant levels with strong or weak versions of the program). More potential outcomes make causal inference more complicated but no different in terms of the basic concepts.

actually exposed to the program: the average treatment effect on the treated, or ATT:

$$\text{ATT} = \mathrm{E}[\mathrm{Y}_i(1) - Y_i(0)|D_i = 1] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]. \qquad 3.$$

The ATT will not equal the ATE if the units exposed to the program respond differently to the program than the unexposed units would respond (i.e., heterogeneous responses), such as if the locations exposed to an antipollution program are more responsive to the program, on average, than locations not exposed to the program would have been had they been exposed. To estimate the ATT, one needs only to seek units for which $E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1]$: i.e., a group of untreated units that represent what would have happened among treated units had they not been treated.

In addition to the ATE and ATT, other causal effects may be of interest, including the average treatment effect on the untreated, or ATU (the expected impact on a randomly chosen untreated unit had it been exposed to the program); local ATEs; intent-to-treat effects; and others. In fact, for any particular outcome and data set, one might estimate myriad causal effects, some of which are policy relevant and others that are not, but each requires different untestable assumptions to estimate them from the data (7). Without additional modifiers, terms such as impact and causal effect shed little light on the policy relevance of an estimated impact or on the assumptions required to draw inferences about its value from data.

Rival Explanations

Every scientist has heard the adage that correlation does not imply causation. The adage has a kernel of truth but is often erroneously interpreted as implying that one cannot draw causal inferences from correlational data or that all correlations have equal value (or equally lack value). The adage's kernel of truth is captured in **Figure 1**. We wish to estimate a causal relationship between an intervention, D, and an outcome, Y, and believe that D affects Y through some mechanism(s), M. A correlation between values of D and Y may not reflect a causal relationship between D and Y, just as the absence of a correlation would not necessarily reflect the absence of a causal relationship.

Drawing causal inferences from correlations is complicated by the presence of confounders: observable variables, X, and unobservable variables, U, which are common causes of both D and Y.

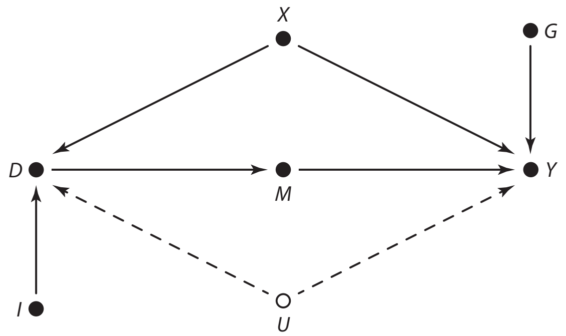

Figure 1

A directed acyclic graph in which a single-headed arrow (→) represents a causal link or pathway between two variables, represented by nodes. Four empirical designs for estimating causal effect of treatment (D) on outcome (Y) are depicted: (1) Exert experimental control on D; (2) condition on observable confounders (X) and assume that unobserved confounders (represented by the dashed singled headed arrows and open node at U) do not exist; (3) exploit variation in D that comes from a source (I) that has no path to Y except through D; and (4) identify an isolated and exhaustive set of mechanisms (M).

Any correlation between D and Y may thus reflect the influence of confounders rather than a true causal relationship between D and Y. In other words, the assignment of D may not be independent of potential outcomes, and we thus have to be careful about using outcomes we can observe as surrogates for counterfactual outcomes.

For example, until recently, the conservation literature measured the environmental and social impacts of protected areas, such as national parks, by contrasting outcomes in protected and unprotected areas. The average outcome in unprotected areas ($E[Y_i(0) \mid D_i = 0]$) served as a surrogate for the average counterfactual outcome of the protected areas had they not been protected ($E[Y_i(0) \mid D_i = 1]$), as if these areas were randomly assigned. Protected areas, however, follow a well-known global pattern of assignment: They are systematically placed in areas that are less productive for alternative economic uses (8–10).

If we were to ignore this selection process and simply estimate the correlation between protection and an outcome, we could not determine if the correlation arose from a causal effect of protection or, for example, from the differences in productive characteristics between protected and unprotected areas. In other words, if variables X and U exist, but are ignored by the analyst, the estimator of the correlation between D and Y contains hidden bias (in the social sciences, the source of hidden bias is often termed unobservable heterogeneity and the problem it creates is called endogeneity). Such bias is different from sampling error (i.e., a correlation exists between D and Y in a sample by chance); tests developed to characterize the uncertainty that arises from sampling variability (e.g., t-tests, chi-squared tests) do not characterize the uncertainty associated with hidden bias.

The variables X and U are rival explanations for the observed relationship between D and Y. To confidently estimate the impacts of environmental interventions, scientists must seek to eliminate such rival explanations. The quality of an impact evaluation is a function of how well one identifies and eliminates rival explanations and, when one cannot eliminate rival explanations, how well one considers the implications of the rival explanations (see below). The search to eliminate rival explanations puts empirical design, not empirical methods, front and center.

All designs aimed at causal inference can be put into four categories, although the specific causal effects that a design can identify within the same population can vary:

1. Designs that depend on experimental control in which D is varied in ways that are unrelated to potential outcomes [e.g., randomized controlled trials (RCTs)]. Rival explanations are thus eliminated by the experimental control.
2. Designs that depend on conditioning strategies (e.g., regression, matching, fixed effects estimators) in which one assumes that hidden bias comes only from observable variables (X) or from unobservable variables that are either perfectly correlated with observable variables or are time invariant. By controlling for, or blocking, these relationships between D and Y, one eliminates them as rival explanations [called the "back-door criterion" by Pearl (1)].
3. Designs that depend on naturally occurring sources of variation in D that are unrelated to potential outcomes (e.g., instrumental variables and discontinuities). In these designs, an observable variable, I, is assumed to be unrelated to potential outcomes except through its effect on D. These designs are often called natural experiments.
4. Designs that depend on identifying a set of mechanisms, M, through which the effect of D and Y can be estimated through a two-step process that estimates the effect of D on M without bias, and then estimates the effect of M on Y without bias [called the "front-door criterion" by Pearl (1)].

These designs highlight a fundamental feature of all good impact evaluations: Understanding selection is absolutely critical. By selection we mean the process by which some units come to be

“THE EFFECTS OF CAUSES” VERSUS “THE CAUSES OF EFFECTS”

Figure 1 illustrates two important insights from recent advances in causal inference. First, to develop an unbiased estimator of an effect of D on Y (the effect of a cause), we do not need to know G, which captures all the other factors that influence Y. We need focus only on potential sources of hidden bias: X and U. Second, tackling hidden bias in a study that aims to estimate the effects of a single cause is difficult. Tackling such bias in a study that aims to estimate the effects of multiple causes (G) on a single Y (the causes of an effect) is, in our opinion, beyond the reach of current theory and data. For example, a study that purports to estimate the causes (determinants) of deforestation would be better viewed as a study that generates hypotheses for future studies of individual causes, rather than a study that credibly estimates causal relationships between myriad variables and deforestation. The assumptions required to identify the effect of a single cause can be daunting; the likelihood that the same assumptions hold for many variables in a single study is low.

exposed to the treatment and others are not. In the absence of a solid understanding of selection, credible causal inference is simply not possible (see sidebar, "‘The Effects of Causes’ Versus ‘The Causes of Effects’").

Thus, although impact evaluation is an empirical effort, a strong, elaborate theory of the causal pathways and the potential confounders is absolutely critical. Such theories, combined with a deep understanding of field conditions, elucidate the potential designs that can be used to estimate an impact and highlight ways of probing the untestable assumptions required to assign a causal interpretation to a correlation. They also help guide efforts to elucidate the heterogeneity of impacts and the causal mechanisms (see section on Heterogeneity and Mechanisms).

EXPERIMENTAL DESIGNS

Randomization and Experimental Control

Experimental control, particularly through randomization, makes the estimation of counterfactual outcomes easier. An intervention assigned at random is assigned independent of potential outcomes (i.e., in **Figure 1**, there are no X or U confounders). The treated and untreated groups are just two random samples from the population. The expected potential outcomes of the two groups are thus equal:

$$E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0] = E[Y_i(0)] \tag{4.}$$

$$E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1] = E[Y_i(1)]. \tag{5.}$$

We can therefore use untreated group outcomes as a surrogate for the counterfactual outcomes of the treated group and treated group outcomes as a surrogate for the counterfactual outcomes of the untreated group. With experimental control, the main threat to inference is sampling variability, not bias, and we have well-developed methods for characterizing the uncertainty associated with sampling variability (e.g., t-tests).[2]

[2] All designs have two implicit assumptions: (*a*) common support and (*b*) no interference among units. We do not discuss them because of space constraints, but they are important. Common support means that we can define a causal effect only for units that can be potentially exposed to treatment and control conditions. No interference among units—also known as the stable unit treatment value assumption (SUTVA)—means that whether or not unit i is treated has no effect on unit j’s potential

Despite its advantages, experimental design is rare in the context of environmental programs (15). Yet as scholars and practitioners recognize its potential and see its application to nonenvironmental programs, it is increasingly being applied. Most experimental designs have been applied to test new programs or program elements in the energy and water domains (e.g., 16–21). Other experiments test different ways of implementing existing programs (e.g., 22, 23).

In some cases, randomizing interventions to a single unit (e.g., household) is not possible or desirable, but randomizing interventions to groups of units (e.g., villages) is possible. Clustered RCTs are common in other fields (e.g., education) and are being tried in the context of environmental policy. For example, in Uganda, the United Nations Environmental Program has developed a project aimed at assessing the impact of payments for ecosystem services on poverty and deforestation. Communities were selected at random for the intervention from a pool of communities considered at risk for deforestation (15).

Other Experimental Designs

Experimental control can also be applied in other ways. For example, in a home weatherization program designed to reduce energy use, no household was excluded from participation, but some households were randomly given incentives to participate (24). Thus, some households were exposed to the program for reasons unrelated to their potential energy use. Randomized encouragement designs identify only the impact for the subpopulation of units whose decision to participate is affected by the encouragement, a so-called local average treatment effect (LATE). Units that, despite the encouragement, fail to participate in the program, or units that would have participated even without the encouragement, may respond differently. A LATE may be policy relevant (i.e., the scaled-up program will include encouragement) or it may be better than having no estimate of impact at all (25). To map the LATE to broader impacts like the ATE or the ATT, more assumptions, based on theory and field knowledge, are required (26).

Experimental designs can also provide evidence about the mechanisms through which interventions affect outcomes (27). In such designs, the mechanisms themselves, rather than the interventions, are randomized. For example, one might hypothesize that some protected areas are more effective than others because they have clear boundary demarcation and local community participation. One could randomize these mechanisms in space or time to draw inferences about their contribution to environmental impacts.

Gold Standard?

Experimental designs are neither feasible nor desirable for all contexts [for conditions under which such designs are most desirable and ethical, see Ferraro (15)]. We do not wish to raise experimental designs on a pedestal and assert that they are the gold standard. In practice, they are subject to many threats to validity, including noncompliance, attrition, and randomization biases (26), and are justly criticized for their frequent lack of external validity and inability to elucidate mechanisms. Although there have been important advances in addressing these critiques (26), the role of experimental design in environmental and resource policy evaluations will be modest.

However, when done well, experimental designs are open to fewer challenges than nonexperimental designs. Moreover, even if randomization of an intervention were impossible, the RCT is

outcomes (i.e., no spillovers). In the context of the environment and natural resources, SUTVA is not trivial, but advances have been made in addressing violations of this assumption (e.g., 11), and studies have illustrated empirical methods to test for such violations (e.g., 12–14).

an important conceptual benchmark to which all other designs are compared. As described in the next sections, understanding how one's design differs from the idealized RCT benchmark helps clarify the necessary assumptions as well as the particular causal effect that is being estimated.

CONDITIONING DESIGNS

Confounders

In the environmental literature, researchers typically seek to eliminate rival explanations through conditioning designs. The most common conditioning design asserts that the confounders are observable (X in **Figure 1**) and that one can eliminate their influence through methods such as regression or matching. Although the methods differ, their underlying assumptions are the same. To estimate the ATE, one makes the following assumptions:

$$\text{Assumption 1: } E[Y_i(0)|D_i = 1, X] = E[Y_i(0)|D_i = 0, X] = E[Y_i(0)|X], \tag{6.}$$

$$\text{Assumption 2: } E[Y_i(1)|D_i = 0, X] = E[Y_i(1)|D_i = 1, X] = E[Y_i(1)|X]. \tag{7.}$$

In other words, once we condition on X, valid counterfactuals for treated and untreated groups can be estimated from observable data, on average.[3] Unobservable confounders like U are assumed not to exist, or are perfectly correlated with X.

For example, assume that sex is a confounder. Males are much more likely to be exposed to D and are more likely to have high values of Y. A simple comparison of average outcomes for treated and untreated individuals will reflect the influence of sex, as well as any causal relationship between D and Y. Instead, we could compare average outcomes separately within the subgroups of males and females. Within each of these subgroups, sex cannot be a confounder. To obtain an estimate of the ATE, we invoke Assumptions 1 and 2 and take a weighted average of the two subgroup estimates (weighted by percentages of males and females in the sample). If we are interested only in the ATT, we need invoke only Assumption 1 and use a different weighting (by percentages of males and females in the treated group). In other words, we need only assume that, after conditioning on X (sex), the treated and untreated groups would have the same mean outcomes in the absence of D. In the presence of D, the treated and untreated groups could respond differently, on average, to D, even after conditioning on X (Assumption 2 can be violated).

The dilemma for researchers is determining what variables comprise X. To guide their efforts, researchers need strong, elaborate theories and a solid understanding of the program and field conditions. In any applied context, however, there will be limits to how many potential confounding variables can be measured and included. The most important variables are those that are known to have a strong influence on both treatment selection and outcome, followed by variables that have a strong effect on selection and a weak to moderate effect on outcomes. These variables should be measured before the program starts. Conditioning on postprogram values can introduce bias if the variables are affected, directly or indirectly, by the program. Design-replication studies that contrast the results from a conditioning design with the results from an experimental design benchmark using the same treated group imply that having preprogram outcome values is important; without them, impact estimates from conditioning designs are less likely to be accurate

[3] Some scholars prefer to invoke a stronger version of these two conditional mean independence assumptions, often called ignorability or the conditional independence assumption (CIA). The CIA implies that, after conditioning on X, D is independent of $Y(0)$ and $Y(1)$ and all transformations of $Y(0)$ and $Y(1)$, i.e., $(Y(1),Y(0)) \perp D|X$. In words, once we condition on X, exposure to the program is "as if" randomly assigned.

(28–32). These studies also imply that drawing control and treated units from similar economic and biophysical environments is also important. The task of identifying potential confounders, and collecting accurate data on them, is far more important than the task of identifying the methods through which one will condition on X.

Matching and Regression Methods

The most common methods for conditioning are regression and matching. Contrasting the two methods is beyond the scope of this review (see 4, 5). Given the paucity of matching studies in the environmental literature, we focus on the use of matching methods and emphasize a few important contrasts with regression. Readers interested in more details can find good summaries of both common matching methods (33, 34) and recently developed matching methods (35, 36).

Although many algorithms exist for matching treated and untreated units, the intuition behind them is the same. In an RCT, the expected characteristics of the units in the treated and untreated groups are similar. Thus, their expected potential outcomes are similar. Matching is an attempt to reweight the treated and untreated groups in a nonexperimental study so that the two groups look as if they came from an RCT. Specifically, we wish to see that the distribution of the X variables (the assumed confounders) are similar in our treated and our untreated groups (i.e., X can no longer be a rival explanation).

The quality of a matching exercise is judged by both the quality of this covariate balance and the credibility of the conditional mean independence assumptions. No study using matching is complete without evidence that the distributions of the identified confounders are balanced across treated and untreated groups. In some cases, covariate balance can be achieved only for a subset of the relevant population through methods that eliminate poor matches (e.g., calipers or propensity score trimming). In this case, the causal effect being estimated may differ from the effect originally sought.

Although there is no agreement in the literature on exactly how best to judge covariate balance (see 38), **Figure 2** illustrates some popular metrics. Andam et al. (37) wished to estimate the average effect of legal protection on poverty near Costa Rican protected areas (the ATT). However, the distributions of the preprotection characteristics of protected and unprotected communities are quite different. This imbalance is a concern because these same characteristics are also likely to affect poverty in the absence of protection. After matching, however, the imbalance is small. If these seven measures capture the most important confounders, then Assumption 1 is a credible approximation of reality, and a contrast of the average outcomes between protected and matched unprotected communities is an unbiased estimator of the ATT.

Matching methods have been applied to the evaluation of a wide range of environmental policies, including air-quality regulation (39, 40), voluntary pollution standards (41), farmland conservation (42), eco-certification (43), integrated conservation and development projects (44), payments for ecosystem services (14, 45), and ecosystem conservation (46–51).

Like matching, regression methods require Assumptions 1 and 2 to draw causal inferences from regression coefficients (expressed as "the error term of the regression model is uncorrelated with the treatment variable," or "$E[u \mid D, X] = E[u \mid X]$"). However, in finite samples, regression and matching can generate different impact estimates.

Matching has some advantages over regression that may favor it in applied contexts: (*a*) its design and results are more easily communicated to scholars and practitioners who lack strong statistical training; (*b*) it forces researchers to determine if each treated unit has a valid control (i.e., common support)—regression, which depends on extrapolation based on a prespecified model, fails to alert researchers to the incompatibility of treated and control units; (*c*) it is less sensitive to

Covariate balance table

Covariate	Sample	Mean protected	Mean unprotected	Difference in means	Normalized difference	Mean eQQ difference	% Improve mean difference
Preprotection poverty index	Unmatched	13.640	6.573	7.068	0.602	7.079	
	Matched	13.506	13.824	−0.317	0.025	0.495	95.51%
Preprotection % forest	Unmatched	0.512	0.138	0.374	0.577	0.373	
	Matched	0.506	0.505	0.00039	0.001	0.010	99.89%
% Land use capacity 1,2,3	Unmatched	0.078	0.345	−0.267	0.384	0.267	
	Matched	0.077	0.10	−0.023	0.049	0.023	91.36%
% Land use capacity 4	Unmatched	0.170	0.413	−0.243	0.327	0.243	
	Matched	0.170	0.170	−0.00001	0.000	0.013	99.99%
% Land use capacity 5,6,7	Unmatched	0.222	0.193	0.0288	0.043	0.089	
	Matched	0.219	0.206	0.0137	0.023	0.020	52.35%
Distance to major city	Unmatched	58.048	35.490	22.558	0.268	22.510	
	Matched	57.449	55.895	1.553	0.015	5.182	93.11%
Preprotection road networks	Unmatched	441.986	22.837	419.148	0.285	416.354	
	Matched	436.564	281.715	154.849	0.090	157.296	63.05%

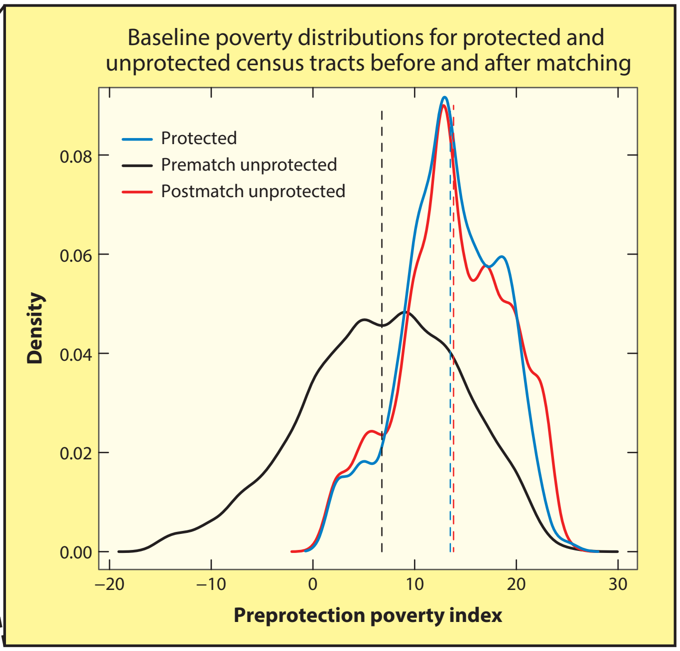

Figure 2

Matching methods seek to achieve covariate balance. Based on data from a study of the impact (ATT) of protected areas on poverty (32), the table shows three popular metrics for assessing covariate balance among protected and unprotected census tracts before and after matching: difference in means, normalized difference in means (the absolute value of the difference in means divided by the square root of the sum of variances of the treated and control groups), and the mean difference in the empirical quantile-quantile (eQQ) plot. As the bottom image depicts, effective matching makes the covariate distributions of protected and unprotected census tracts more similar (dashed vertical lines are means). (ATT, average treatment effect on the treated.)

IMPACT EVALUATION AND THE "QUANTITATIVE VERSUS QUALITATIVE" DEBATE

Social scientists have long debated the contributions of quantitative and qualitative analyses. We are hesitant to enter the fray, but wish to mention four key roles that qualitative information plays in impact evaluations: (*a*) eliminating rival explanations; (*b*) understanding an evaluation's limitations (i.e., external validity); (*c*) identifying causal hypotheses to test; and (*d*) identifying factors that may serve as moderators and mechanisms. With regard to eliminating rival explanations, one should accept support from wherever one can find it. For example, one may observe a positive correlation between variable *D* and variable *Y* and be concerned that the correlation exists because of some other variable, *X*, that may be correlated with *D* and *Y*. One might have information from semistructured interviews in which people describe the key characteristics of a person who is exposed to *D*. If no respondents ever mention *X*, this qualitative information supports eliminating the rival explanation *X*. Nevertheless, we believe it would be hard (but not impossible) to establish a credible claim of causality using only qualitative data, whereas we believe one can establish a credible claim with only quantitative data.

the specific functional form of the relationship between *X* and *Y* (52), which can be a problem for regression when there is severe covariate imbalance between the treated and untreated groups and strong heterogeneity of the causal effects; and (*d*) unlike regression, it permits one to search for covariate balance without ever looking at the outcomes, thereby making research design choices independent of the estimated impact. Recent studies have argued that combining matching and regression may be superior to relying on either alone (32, 34, 52, 53).

Panel Data and Synthetic Control Designs

With additional untestable assumptions and repeated observations of the same units (panel data) before and after treatment assignment, one can also eliminate the effects of unit-specific, unobserved confounders that do not change over the study period (4). Such fixed effects designs are relatively rare in the environmental literature (see, e.g., 54–57). With longitudinal data, researchers more often use so-called random effects models (often called hierarchical models). These models have more statistical power than fixed effects models, but they are much more vulnerable to hidden bias because they assume all sources of confounding are in the model.

Repeated observations can also be useful when the treatment group is a single unit (e.g., a protected area or a watershed) and there are many potential control units. In such cases, a combination of untreated units often better reproduces the characteristics of the treated unit than a single untreated unit alone could (58). This design, which uses transparent rules for creating a synthetic control, can improve the comparative case-control designs that are often used in impact studies of marine protected areas and community natural resource management. We know of no examples of this design in the environmental literature.

INSTRUMENTAL VARIABLE AND DISCONTINUITY DESIGNS

If one believes that there are unobserved sources of bias (*U* in **Figure 1**), one may be able to use theory and a deep knowledge of an intervention's implementation to identify sources of variation in exposure to the intervention that are uncorrelated with potential outcomes. These designs are most commonly implemented through the exploitation of instrumental variables (IVs) and discontinuities in program eligibility (see 4, 5). Such designs are rare in the environmental literature, a reflection of the inchoate state of environmental program evaluation.

Instrumental Variables

An IV is something that affects which units are exposed to the program (selection) but does not affect the outcome except through its effect on the program—i.e., variable *I* in **Figure 1**. For example, Sims (59) uses "priority watershed status" as an IV to measure the effect of protected areas on poverty. She shows that priority status, which is determined by proximity to the headwaters of major rivers, affects the probability that the government will establish a protected area. Then she claims that status is determined by the downstream destination of waterways and that it would therefore neither affect poverty near the protected areas nor be correlated with local characteristics that affect poverty. The negative correlation she finds between priority status and poverty must therefore reflect a causal relationship between protection and poverty reduction. Although a credible IV is hard to find, there are other examples illustrating the use of IV designs to evaluate environmental programs (see, e.g., 60–62).

Like the randomized encouragement designs discussed above, IV designs can only identify a LATE. For example, Sims's (59) IV design estimates the effect of protection on poverty among the localities whose exposure to a protected area is determined by whether or not they are located in the upper watershed. These localities are called compliers. One can sometimes estimate the proportion of the sample that is composed of compliers, but one cannot identify the compliers. A LATE estimate is specific to the IV: A different IV would likely identify a different LATE for the same *D* and *Y*, which can make comparisons across IV studies difficult. Transforming a LATE into broader impacts like the ATE or the ATT requires assumptions, based in theory and field knowledge, that relate the compliers to the rest of the sample (4).

Discontinuities

Like IV designs, discontinuity designs exploit knowledge about the factors that affect program exposure (see, e.g., 54, 63, 64). For example, Alix-Garcia et al. (65) sought to estimate the effect of a Mexican antipoverty program on deforestation. However, many unobservable characteristics that affect which communities are exposed to the program also affect land use, and thus deforestation. To eliminate these rival explanations, the authors took advantage of a program eligibility rule: Only communities with a poverty score over a certain value are eligible for the program. The rule creates a discontinuity in the probability of program participation: It jumps from very low to very high near the threshold score. Given the way the poverty score is calculated, the authors argue that there is no reason to believe there would be a discontinuous jump in deforestation at the threshold in the absence of the program—the communities on either side of the threshold are quite similar. Yet they detect a discontinuous jump in deforestation right at the threshold and conclude that the antipoverty program nearly doubles the probability of deforestation. Using theory and other data, they strengthen these conclusions by showing that jumps do not occur at other values of the score and by identifying and estimating the effects of plausible mechanisms through which the program could affect deforestation. Like IV designs, discontinuity designs identify a treatment effect for a subgroup of units—those that are near the discontinuity. To extrapolate the effect to the broader population, more assumptions and modeling are necessary.[4]

[4] Discontinuities formed by geography, such as a line on a map, are also sometimes used (66). Such designs, however, assume that nothing that affects outcomes changes dramatically at the discontinuity, other than exposure to the program. That assumption can be hard to justify (67).

MEASUREMENT ERROR

Evaluators often ignore the implications of measurement error in their treatment variable, in their outcome variable, and in their conditioning (control) variables. Recent research has demonstrated, however, that these errors are often not random, and ignoring them can lead to serious bias (62, 73). Even worse, it can be quite difficult to sign the direction of the bias in many cases.

INTERROGATING THE ASSUMPTIONS

High-quality evaluations do not end with the estimation of an impact. As noted above, causality cannot be inferred from the data alone; one must also rely on untestable assumptions about counterfactual outcomes. To explore the credibility of these assumptions, evaluators rely on two types of approaches: (*a*) approaches that seek evidence for violations of the assumptions, and (*b*) approaches that explore the implications for the study were such violations to exist. **Figure 3** captures the intuition behind both approaches. In designs that depend on conditioning, we seek evidence of the red line from *U* to *D* (hidden bias) and explore the implications if such a line were to exist. In designs that depend on IVs and discontinuities, we seek evidence for, and explore the implications of, a red line from *U* to *I* or from *I* to *Y*.

Tests of Known Effects

To probe whether the assumption of no hidden bias may be false, analysts can use tests of known effects [also called placebo or falsification tests (2)]. These tests require an elaborate theory of causal pathways and good knowledge of field conditions. In one type of test, the analyst posits the potential unobserved source of bias, *U*. For example, in the case of water conservation incentive programs, one might worry that unobservable proenvironment preferences are stronger among participants than nonparticipants. Next, the analyst identifies an outcome, *Z*, that is known (or

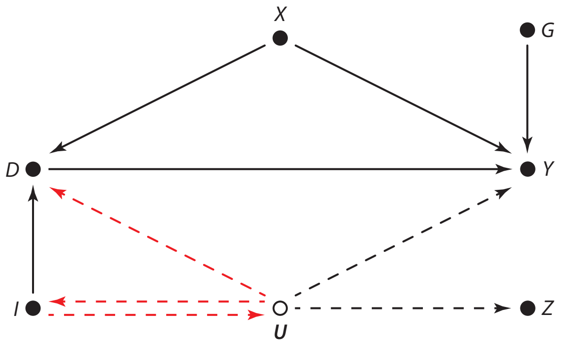

Figure 3

Directed acyclic graph depicts potential sources of hidden bias when estimating an effect of program *D* on outcome *Y* by conditioning on observable confounders *X* or exploiting variation in *D* that comes from variable *I*. These sources are rival causal pathways (*red dashed arrows*) related to unobserved confounders *U*. Tests of known effects attempt to detect these rival pathways by testing for a correlation between *D* or *I* and an outcome, *Z*, that is known (asserted) to be unrelated to *D* or *I* (i.e., theory and experience imply that no arrow goes directly from *D* or *I* to *Z*). A detected correlation implies the presence of a rival pathway through *U*.

INTERNAL VERSUS EXTERNAL VALIDITY

This review focuses on internal validity—whether a correlation implies a causal relationship. In contrast, external validity considers whether an implied causal relationship can be generalized to other settings. Debates about which type of validity is most important sometimes arise. These debates can often be defused by using environmental programs to test the theories that underlie the design of the programs. For example, suppose an elaborate theory predicts that wider local participation in community land management will lead to four changes in environmental and social outcomes through three mechanisms, whose effects are moderated by two factors. If we test that theory using program X in country Z, we create more generalizable and useful knowledge than if we simply found that program X increased basal stem density and household income (one must, however, confirm that the program can be interpreted as embodying the constructs in the theory, called construct validity). We thus address external validity through theory development and testing, rather than by finding a representative program and site in which we measure impacts. As theories become more elaborate and the evidence for them stronger, our knowledge about how best to design environmental programs will grow.

believed) to be affected by *U*, but not by *D*. In the case of the water conservation incentive program, *Z* might be solid waste recycling. Using the design that purports to identify an impact of incentives (*D*) on water conservation (*Y*), the analyst estimates the effect of *D* on *Z*. If the estimated effect were different from zero, it would raise concerns that the design suffers from hidden bias. In other words, it would imply that the red directed edge from *U* to *D* in **Figure 3** exists.

Two other popular tests of known effects include (*a*) demonstrating that the program has no effect on preprogram outcomes but that it does have an effect on postprogram outcomes, and (*b*) confirming a theoretically predicted dose-response relationship between *D* and *Y* (e.g., decomposing *D* into D^{Strong} and D^{Weak} and showing that the effect of D^{Strong} is larger than the effect of D^{Weak}). All tests of known effects are indirect, however, and depend on the validity of theory and field knowledge. None can disconfirm hidden bias.

Partial Identification and Sensitivity Analyses

High-quality impact evaluations must therefore consider the implications of potential hidden bias. In other words, one should posit that the red lines in **Figure 3** may exist and then ask how the conclusions would change (2, 68–70). To do this, studies rely on three approaches: (*a*) an informal approach that merely tries to sign the direction of potential bias (i.e., given the most likely rival explanations, is our impact estimate likely to be an upper or a lower bound on the true impact?); (*b*) partial identification, which starts with weak assumptions about the data-generating process in order to bound the potential values of the impacts, and then progressively makes the assumptions stronger to shrink the plausible bounds (see, e.g., 45); and (*c*) sensitivity analysis, which starts with strong assumptions and then asks how strong the hidden bias would have to be in order to change our conclusions (see, e.g., 71, 72).[5] Like the tests of known effects, all three approaches require a good grasp of the relevant theory and field conditions and none can disconfirm hidden bias.

[5] When multiple control groups are available (e.g., two different surveys on nonparticipants exist), they can be used both to test for hidden bias and to bound the range of plausible impact estimates (see 2, 71).

HETEROGENEITY AND MECHANISMS

Moderators

With a credible estimate of an impact, one can explore the factors that moderate the magnitude of the impact and the mechanisms through which the impact arises. A moderator is a variable that is unaffected by the intervention and whose value affects the magnitude of an impact (e.g., preintervention poverty of the participating communities). A mechanism is a variable that is affected by the intervention and in turn affects the outcome (e.g., community forest management provides a stable source of household revenue, which then reduces poverty).

Figure 4 illustrates two ways to think about moderators. In **Figure 4*a***, a sample of Costa Rican forest parcels is divided into subgroups: areas with high or low preprotection poverty, and areas with high (steep) or low (flat) slopes. The bar charts portray the estimated subgroup impacts of legal protection on poverty and avoided deforestation (conditional ATTs). These estimated conditional impacts include the contribution of all the other characteristics that might be correlated within a subgroup (for other examples, see References 51, 57).

Figure 4*b* presents an alternative way of thinking about heterogeneous effects (see also the sidebar on "Recommendations for Estimating Heterogeneous Effects"). Using partial linear modeling, the conditional impacts are estimated along the entire range of the observable characteristic, holding constant the other observable characteristics. Other studies use interaction terms in a regression specification to estimate conditional mean impacts, holding constant the other observable characteristics (see, e.g., 14, 23, 59, 76).

If one understands what factors moderate the magnitude of an impact, one can improve program cost-effectiveness and avoid harms. For example, Ferraro & Miranda (77) find that by targeting water conservation efforts only at wealthy home-owning households that historically use a lot of water, a US program could reduce its costs by over 50% but suffer only a 20% reduction in its target of water savings and simultaneously avoid the political conflict that can arise from asking

RECOMMENDATIONS FOR ESTIMATING HETEROGENEOUS EFFECTS

- Prior to analysis, select a small (<5) number of theoretically motivated or policy-relevant subgroups.
- If treatment is randomized, randomize it within the selected subgroups.
- When the null hypothesis of the unconditional treatment effect equal to zero cannot be rejected, be wary of estimating conditional effects unless randomization was done within subgroups. First conduct a test of the null hypothesis that the average conditional effect is equal to zero across all subgroups (79).
- When the hypothesis of the treatment effect equal to zero can be rejected, first do a joint test of the null hypothesis that all subgroup effects are equal to zero (e.g., an F-test; a quantile model combined with a rank preservation assumption; or a test of the null of constant conditional average treatment effect, as in Reference 79).
- If the null hypothesis that all subgroup effects are equal to zero can be rejected, first do a subgroup analysis that does not hold everything else equal before attempting to estimate conditional effects that hold everything else constant (e.g., partial linear modeling).
- For subgroup analyses that rely on multiple hypothesis testing, maintain a constant family-wise type I error rate.
- Interrogate the assumptions.

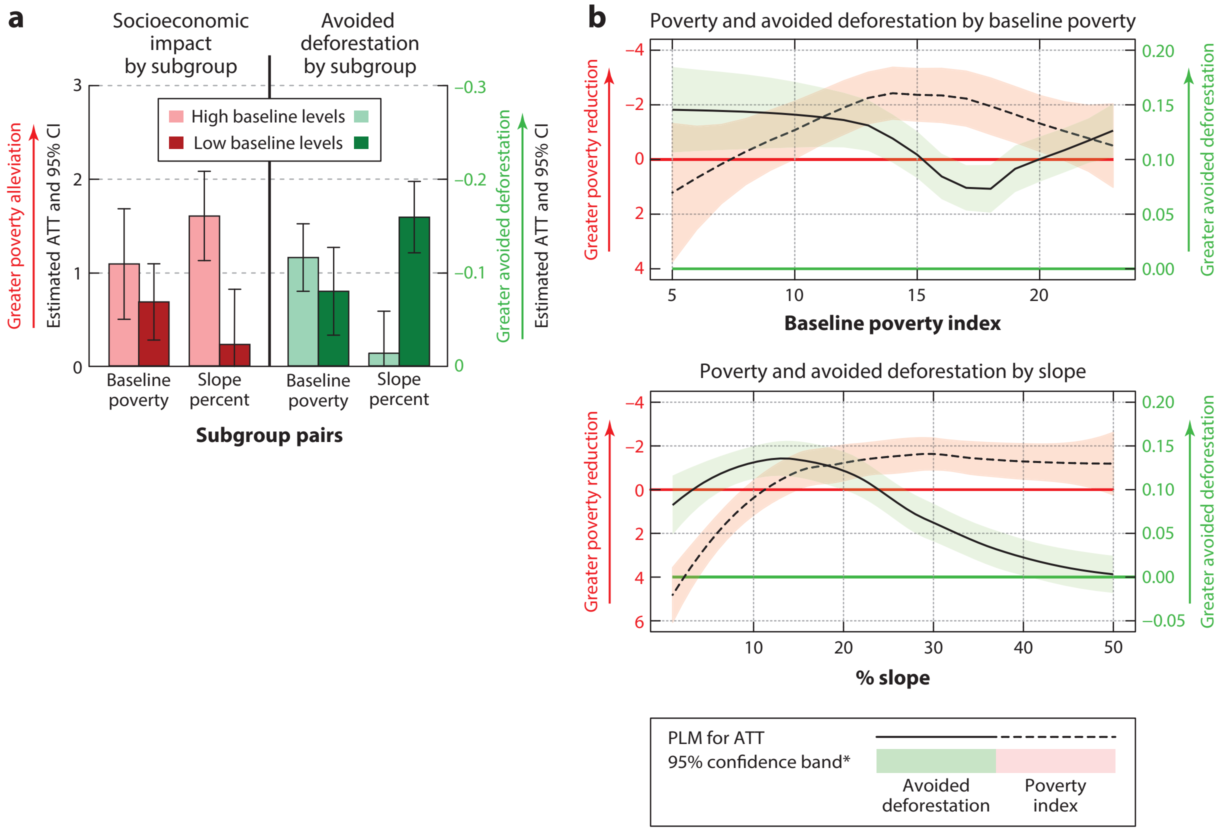

Figure 4

Heterogeneous impacts. Based on data from two studies (74, 75), panel *a* depicts the estimated impacts of Costa Rica's protected areas on deforestation and poverty conditional on high or low values of observable moderators of the impacts (i.e., conditional ATTs). Based on results from a nonparametric partial linear model (PLM), panel *b* depicts estimates of the conditional ATT along the entire range of moderator values, after controlling for the influence of other moderators. (CI, confidence interval; ATT, average treatment effect on the treated.)

low-income users to reduce their water use. Understanding which factors moderate program impacts can also help practitioners predict program impacts in new areas or in the future.

If one understands what factors moderate multiple impacts, one can explore potential tradeoffs in different targeting strategies. For example, Ferraro et al. (75) find that, for protected areas in Costa Rica and Thailand, the targeting rules that would lead to the largest reductions in deforestation are not necessarily the rules that would lead to the largest reductions in poverty. **Figure 4*b*** illustrates their results for Costa Rica.

Estimating conditional causal effects requires even more elaborate theories and more untestable (and often less credible) assumptions than were required to estimate unconditional effects. Thus, even in RCTs, subgroup effects are considered much less credible than unconditional effects (78).

Mechanisms

In contrast to a moderator, a mechanism lies on the causal pathway between *D* and *Y*. **Figure 5** illustrates potential mechanisms for an intervention that creates community forest management

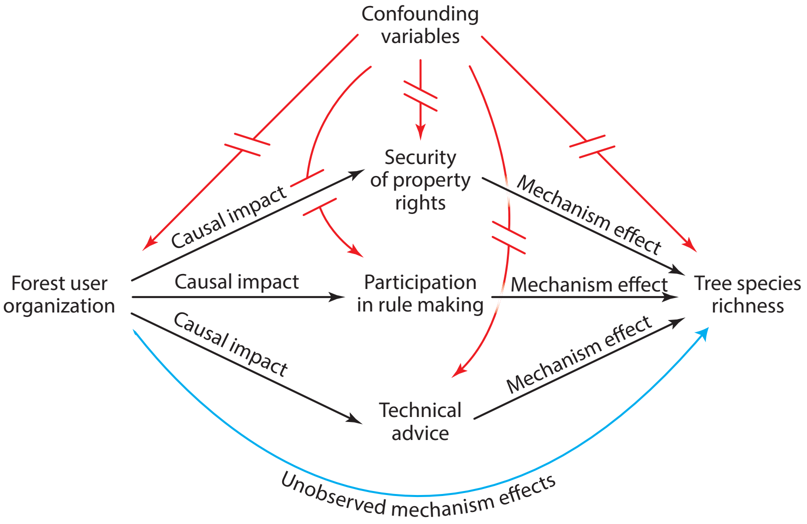

Figure 5

Directed acyclic graph depicts the framework for estimating causal mechanism effects through a two-step process: estimate the impact of treatment (establishment of a forest user organization) on the mechanisms (security of property rights, participation in rule making, and technical advice), and then estimate the impact of the mechanisms on the outcome (tree species richness). When the set of observable mechanisms is not exhaustive, the overall impact can be decomposed into observed mechanism effects and unobserved mechanism effects. The broken, red, single-headed arrows illustrate that conditioning breaks the confounding causal pathways to help identify the causal mechanism effects.

institutions. The contributions of each mechanism to changes in species richness can be estimated in a two-step process. First, one estimates a causal effect of creating institutions on the mechanisms. Second, one estimates how the program's effect on each mechanism affects species richness. **Figure 5** also illustrates the obstacles. First, one must control for confounding variables that jointly affect intervention, mechanisms, and outcome. Second, one must model the effect of the intervention on the outcome in the absence of the mechanisms.

Estimates of mechanism effects are rare in the environmental literature. Using a two-stage approach, Ferraro & Hanauer (80) estimate the individual contributions of mechanisms through which protected areas affected poverty in Costa Rica. Their results imply that two-thirds of the observed reduction in poverty caused by protected areas came from tourism and the other third from unmeasured mechanisms. The effects of protection on land use and on infrastructure had no detectable effect on poverty, on average.

An alternative perspective on mechanisms views them as a means to estimate a causal effect of *D* on *Y* without bias. Although mechanism-based impact evaluations are entirely absent in the environmental literature, they are essentially just a recasting of what we described above. For example, Ferraro & Hanauer's (80) mechanism study assumes that, based on results from a previous study (37), Costa Rica's protected areas did reduce local poverty. They then estimated the relative contributions of three potential mechanisms to this overall impact. An alternative interpretation of their study would question whether the original 2010 study of protected areas' effect on poverty was free from hidden bias. Instead, one could appeal to theory to make the assumption that, were such an effect to exist, it would come through the three mechanisms. The two-stage empirical design would still be used, but the results would be interpreted differently. If one were

EVIDENCE ABOUT THE VALUE OF EVIDENCE

Throughout this review, we implicitly assume that evidence about impacts is important for improving environmental programs. Yet there is little evidence to support this assumption. Scholars and practitioners can better understand how evidence about impacts, and the ways it is communicated, ultimately affects environmental and social outcomes by applying the same designs used to generate evidence about environmental program impacts (see **Figure 1**).

willing to assume that the three mechanisms completely captured all possible mechanisms, then the sum of the mechanism effects is equal to the ATT (if the set of mechanisms were not exhaustive, the sum equals only a fraction of the ATT). In other words, one can either assume that D affects Y, and then estimate the effects of different components of M, or one can assume that the components of M are known, and then use that assumption to estimate the effect of D on Y. As with all causal inference, nothing can be achieved without committing to a set of untestable assumptions.

CONCLUSIONS

We have no shortage of good ideas to solve environmental problems, but we do have a glaring shortage of evidence to support these ideas. To generate evidence, we must deliberately design environmental programs with the aim of estimating their impacts and use appropriate methods to understand cause and effect from the available data. These two actions are the hallmarks of evidence-based policy making. Unfortunately, they are still rare in environmental science and policy. As a result, we have myriad publications filled with data on the status and trends of our biosphere but surprisingly little evidence on the effects of our policies that aim to change the status and trends (see the sidebar on "Evidence About the Value of Evidence").

Producing high-quality evidence will require much more than the application of sophisticated statistical methods. Methods, no matter how sophisticated, cannot substitute for appropriate empirical designs. A focus on design means precisely describing the causal effect to be estimated, the way in which rival explanations will be eliminated, and the untestable assumptions that will be called upon to make causal inferences. Strong, elaborate theories and field knowledge are required to identify appropriate empirical designs, to help identify and eliminate rival explanations, and to consider the implications of potential hidden biases.

A focus on design also implies collecting different (not more) data. Scholars and practitioners invest much of their time and money measuring observable outcomes in myriad ways, but they fail to collect data that helps estimate counterfactual outcomes. Without credible estimates

SYSTEMATIC REVIEWS AND META ANALYSES

As the evidence from credible impact evaluations grows, systematic reviews can be used to synthesize and communicate what is known about the effects of a particular program or class of programs (81). Systematic reviews are different from classic literature reviews in that systematic reviews have clear rules for deciding which studies to include and how much weight (credibility) to assign to each study. The Collaboration for Environmental Evidence (**http://www.environmentalevidence.org**) and the new journal *Environmental Evidence* are dedicated to encouraging wider use of systematic reviews in the environmental domain.

of, or bounds on, counterfactual outcomes, there can be no evidence about impacts. In other fields with stronger evidence bases (e.g., medicine, labor policy, public health, and, more recently, education), scholars and practitioners have developed partnerships that generate relevant data and use credible designs to analyze the data. The environmental community needs to do the same.

Rigorous impact estimates may not be possible for many environmental programs and projects. Scholars and practitioners must collaborate to identify programs where high-quality evaluations are feasible and likely to generate knowledge that benefits other contexts. Nevertheless, if the community of environmental scientists and practitioners continues to believe that money spent measuring impact is money not available for solving problems, it will ensure that we remain unable to distinguish our successes from our failures, and thus unable to make progress.

SUMMARY POINTS

1. The quality of any impact evaluation, regardless of method or data, is a function of how well it identifies and eliminates rival explanations for what is observed.
2. Identifying and eliminating rival explanations requires a solid understanding of selection. Selection refers to the process by which some units come to be exposed to the program and others do not. Without a credible explanation of selection, credible causal inferences are not possible.
3. High-quality impact evaluations
   - acknowledge that not all rival explanations can be eliminated, and explore the implications of these rival explanations for our understanding of impacts;
   - emphasize design over methods;
   - clarify the assumptions required for causal inferences and then interrogate these assumptions with theory and data;
   - require elaborate theories and deep field knowledge; and
   - focus on finding credible ways to estimate counterfactual outcomes.
4. Attempts to estimate the heterogeneity of impacts, and the mechanisms through which impacts arise, require even more attention to theory and design—and warrant even more skepticism—than is required and warranted by estimates of average unconditional impacts.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press
2. Rosenbaum PR. 2002. *Design of Observational Studies*. New York: Springer
3. Rubin DB. 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* 26(1):20–36

4. Angrist JD, Pischke JS. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press
5. Morgan SL, Winship C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge Univ. Press
6. Ferraro PJ. 2009. Counterfactual thinking and impact evaluation in environmental policy. *New Dir. Eval.* 122:75–84
7. Heckman JJ, Vytlacil E. 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73(3):669–738
8. Millenn. Ecosyst. Assess. 2005. *Ecosystems and Human Well-Being: A Framework for Assessment.* Washington, DC: Island
9. Joppa LN, Pfaff A. 2009. High and far: biases in the location of protected areas. *PLOS ONE* 4(12):e8 2739
10. Pressey RL, Bottrill MC. 2008. Opportunism, threats, and the evolution of systematic conservation planning. *Conserv. Biol.* 22(5):1340–45
11. Aronow PM, Samii C. 2013. Estimating average causal effects under interference between units. arXiv:1305.6156
12. Andam KS, Ferraro PJ, Pfaff A, Sanchez-Azofeifa GA, Robalino JA. 2008. Measuring the effectiveness of protected area networks in reducing deforestation. *Proc. Natl. Acad. Sci. USA* 105(42):16089–89
13. Gaveau DLA, Epting J, Lyne O, Linkie M, Kumara I, et al. 2009. Evaluating whether protected areas reduce tropical deforestation in Sumatra. *J. Biogeogr.* 36(11):2165–75
14. Alix-Garcia JM, Shapiro EN, Sims KRE. 2012. Forest conservation and slippage: evidence from Mexico's national payments for ecosystem services program. *Land Econ.* 88(4):613–38
15. Ferraro PJ. 2012. *Experimental project designs in the Global Environment Facility: designing projects to create evidence and catalyze investments to secure global environmental benefits*. STAP Advis. Doc., Glob. Environ. Facil., Washington, DC
16. Ferraro PJ, Miranda JJ, Price MK. 2011. The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *Am. Econ. Rev. Pap. Proc.* 101(3):318–22
17. Allcott H. 2011. Social norms and energy conservation. *J. Public Econ.* 95(9):1082–95
18. Allcott H. 2011. Rethinking real-time electricity pricing. *Resour. Energy Econ.* 33(4):820–42
19. Wolak FA. 2011. Do residential customers respond to hourly prices? Evidence from a dynamic pricing experiment. *Am. Econ. Rev.* 101(3):83–87
20. Jessoe K, Rapson D. 2014. Knowledge is (less) power: experimental evidence from residential energy use. *Am. Econ. Rev.* 104(4):1417–38
21. Ferraro PJ, Price MK. 2013. Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. *Rev. Econ. Stat.* 95(1):64–73
22. Duo E, Greenstone M, Pande R, Ryan R. 2013. Truth-telling by third-party auditors and the response of polluting firms: experimental evidence from India. *Q. J. Econ.* 28(4):1499–545
23. Bennear L, Tarozzi A, Pfaff A, Balasubramanya S, Ahmed M, van Geen A. 2013. Impact of a randomized controlled trial in arsenic risk communication on household water-source choices in Bangladesh. *J. Environ. Econ. Man.* 65(2):225–40
24. Fowlie M, Greenstone M, Wolfram C. 2014. *Reducing energy consumption and greenhouse gas emissions through energy efficient retrofits: evidence from low-income households*. Rep., Poverty Action Lab, Cambridge, MA. **http://www.povertyactionlab.org/evaluation/reducing-energy-consumption-and-greenhouse-gas-emissions-through-energy-efficient-retrofi**
25. Imbens GW. 2010. Better late than nothing: some comments on Deaton 2009 and Heckman and Urzua 2009. *J. Econ. Lit.* 48:399–423
26. Gerber AS, Green DP. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton
27. Ludwig J, Kling JR, Mullainathan S. 2011. Mechanism experiments and policy evaluations. *J. Econ. Perspect.* 25(3):17–38
28. Heckman J, Ichimura H, Todd P. 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Rev. Econ. Stud.* 64(4):605–54

29. Heckman J, Ichimura H, Todd P. 1998. Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* 65(2):261–94
30. Smith J, Todd P. 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? *J. Econom.* 125:305–53
31. Cook T, Shadish W, Wong V. 2008. Three conditions under which observational studies produce the same results as experiments. *J. Policy Anal. Manag.* 274:724–50
32. Ferraro PJ, Miranda JJ. 2014. The performance of non-experimental designs in the evaluation of environmental policy: a design-replication study using a large-scale randomized experiment as a benchmark. *J. Econ. Behav. Organ.* In press
33. Stuart EA. 2010. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1):1
34. Imbens GW, Wooldridge JM. 2009. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47(1):5–86
35. Diamond A, Sekhon JS. 2013. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* 95(3):932–45
36. Hainmueller J. 2012. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* 20(1):25–46
37. Andam KS, Ferraro PJ, Sims KRE, Healy A, Holland MB. 2010. Protected areas reduced poverty in Costa Rica and Thailand. *Proc. Natl. Acad. Sci. USA* 107(22):9996–10001
38. Lee WS. 2013. Propensity score matching and variations on the balancing test. *Empir. Econ.* 44(1):47–80
39. List JA, Millimet DL, Fredriksson PG, McHone WW. 2003. Effects of environmental regulations on manufacturing plant births: evidence from a propensity score matching estimator. *Rev. Econ. Stat.* 85(4):944–52
40. Greenstone M. 2004. Did the clean air act cause the remarkable decline in sulfur dioxide concentrations? *J. Environ. Econ. Manag.* 47(3):585–611
41. Blackman A, Rivera J. 2010. *The evidence base for environmental and socioeconomic impacts of sustainable certification*. RFF Discuss. Pap., Resour. Futur., Washington, DC. **http://www.rff.org/RFF/Documents/EfD-DP-10-10.pdf**
42. Mezzatesta M, Newburn DA, Woodward RT. 2013. Additionality and the adoption of farm conservation practices. *Land Econ.* 89(4):722–42
43. Blackman A, Naranjo A. 2012. Does eco-certification have environmental benefits? Organic coffee in Costa Rica. *Ecol. Econ.* 83:58–66
44. Weber JG, Sills EO, Bauch S, Pattanayak SK. 2011. Do ICDPs work? An empirical evaluation of forest-based microenterprises in the Brazilian Amazon. *Land Econ.* 87(4):661–81
45. Arriagada RA, Ferraro PJ, Sills EO, Pattanayak SK, Cordero-Sancho S. 2012. Do payments for environmental services affect forest cover? A farm-level evaluation from Costa Rica. *Land Econ.* 88(2):382–99
46. Joppa L, Pfaff A. 2012. Reassessing the forest impacts of protection: the challenge of nonrandom location and a corrective measure. *Ann. N. Y. Acad. Sci.* 1185:135–39
47. Honey-Roses J, Baylis K, Ramírez MI. 2011. A spatially explicit estimate of avoided forest loss. *Conserv. Biol.* 25(5):1032–43
48. Gaveau DLA, Curran LM, Paoli GD, Carlson KM, Wells P, et al. 2012. Examining protected area effectiveness in Sumatra: importance of regulations governing unprotected lands. *Conserv. Lett.* 5(2):142–48
49. Andam KS, Ferraro PJ, Hanauer MM. 2013. The effects of protected area systems on ecosystem restoration: a quasi-experimental design to estimate the impact of Costa Rica's protected area system on forest regrowth. *Conserv. Lett.* 6(5):317–23
50. Ferraro PJ, Hanauer MM, Miteva DA, Canavire-Bacarezza G, Pattanayak SK, Sims KRE. 2013. More strictly protected areas are not necessarily more protective: evidence from Bolivia, Costa Rica, Indonesia, and Thailand. *Environ. Res. Lett.* 8:025011
51. Nolte C, Agrawal A. 2013. Linking management effectiveness indicators to observed effects of protected areas on fire occurrence in the Amazon rainforest. *Conserv. Biol.* 27(1):155–65

52. Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* 15:199–236
53. Abadie A, Imbens GW. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–67
54. Bennear LS, Olmstead SM. 2008. The impacts of the right to know: information disclosure and the violation of drinking water standards. *J. Environ. Econ. Manag.* 56(2):117–30
55. Sims KRE, Schuetz J. 2009. Local regulation and land-use change: the effects of wetlands bylaws in Massachusetts. *Reg. Sci. Urban Econ.* 39(4):409–21
56. Burgess R, Hansen M, Olken BA, Potapov P, Sieber S. 2012. The political economy of deforestation in the tropics. *Q. J. Econ.* 127(4):1707–54
57. Davis L, Fuchs A, Gertler P. 2014. Cash for coolers: evaluating a large-scale appliance replacement program in Mexico. *Am. Econ. J. Econ. Policy*. In press
58. Abadie A, Diamond A, Hainmueller J. 2010. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J. Am. Stat. Assoc.* 105(490):493–505
59. Sims KRE. 2010. Conservation and development: evidence from Thai protected areas. *J. Environ. Econ. Manag.* 60(2):94–114
60. Chay K, Dobkin C, Greenstone M. 2003. The Clean Air Act of 1970 and adult mortality. *J. Risk Uncertain.* 27(3):279–300
61. Busch J, Cullen R. 2009. Effectiveness and cost-effectiveness of yellow-eyed penguin recovery. *Ecol. Econ.* 68(3):762–76
62. Millimet DL. 2011. *The Elephant in the Corner: A Cautionary Tale About Measurement Error in Treatment Effects Models*, Vol. 27. Bingley, UK: Emerald Group
63. Chay K, Greenstone M. 2005. Does air quality matter? Evidence from the housing market. *J. Polit. Econ.* 113(2):376–424
64. Cutter WB, Neidell M. 2009. Voluntary information programs and environmental regulation: evidence from spare the air. *J. Environ. Econ. Manag.* 58(3):253–65
65. Alix-Garcia J, McIntosh C, Sims KRE, Welch JR. 2013. The ecological footprint of poverty alleviation: evidence from Mexico's Oportunidades program. *Rev. Econ. Stat.* 95(2):417–35
66. Liscow ZD. 2013. Do property rights promote investment but cause deforestation? quasi-experimental evidence from Nicaragua. *J. Environ. Econ. Manag.* 65(2):241–61
67. Deaton A. 2010. Instruments, randomization and learning about development. *J. Econ. Lit.* 48:424–55
68. Altonji J, Elder TE, Taber CR. 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic school. *J. Polit. Econ.* 105:151–84
69. Ichino A, Mealli F, Nannicini T. 2008. From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *J. Appl. Econom.* 23(3):305–27
70. Manski CF. 2011. Policy analysis with incredible certitude. *Econ. J.* 121(554):F261–89
71. Canavire-Bacarezza G, Hanauer MM. 2013. Estimating the impacts of Bolivia's protected areas on poverty. *World Dev.* 41:265–85
72. Ferraro PJ, McIntosh C, Ospina M. 2007. The effectiveness of the US Endangered Species Act: an econometric analysis using matching methods. *J. Environ. Econ. Manag.* 54(3):245–61
73. Butler JS, Moser C. 2007. Cloud cover and satellite images of deforestation. *Land Econ.* 83(2):166–73
74. Ferraro PJ, Hanauer MM. 2011. Protecting ecosystems and alleviating poverty with parks and reserves: 'win-win' or tradeoffs? *Environ. Resour. Econ.* 48(2):269–86
75. Ferraro PJ, Hanauer MM, Sims KRE. 2011. Conditions associated with protected area success in conservation and poverty reduction. *Proc. Natl. Acad. Sci. USA* 108(34):13913–18
76. Pfaff A, Robalino J, Sanchez-Azofeifa GA, Andam KS, Ferraro PJ. 2009. Park location affects forest protection: land characteristics cause differences in park impacts across Costa Rica. *B.E. J. Econ. Anal. Policy* 9(2):5
77. Ferraro PJ, Miranda JJ. 2013. Heterogeneous treatment effects and causal mechanisms in non-pecuniary, information-based environmental policies: evidence from a large-scale field experiment. *Res. Energy Econ.* 35:356–79
78. Sun X, Briel M, Busse JW, You JJ, Akl EA, et al. 2012. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 344:e1553

79. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2008. Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* 90(3):389–405

80. Ferraro PJ, Hanauer MM. 2014. Quantifying causal mechanisms to determine how protected areas affect poverty through changes in ecosystem services and infrastructure. *Proc. Natl. Acad. Sci. USA* 111(11):4332–37

81. Pullin AS, Knight T. 2013. Time to build capacity for evidence synthesis in environmental management. *Environ. Evid.* 2(1):21