SPECIAL FEATURE: PERSPECTIVE

On design-based empirical research and its interpretation and ethics in sustainability science

Christopher B. Barrett[a,1]

Edited by Paul J. Ferraro, Carey Business School and Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD, and accepted by Editorial Board Member Arun Agrawal June 3, 2021 (received for review February 9, 2021)

Generating credible answers to key policy questions is crucial but difficult in most coupled human and natural systems because complex feedback mechanisms can confound identification of the causal mechanisms behind observed phenomena. By using explicit research designs intended to isolate the causal effects of specific interventions on community monitoring of common property resources, and on the well-being of those resources and their human neighbors, the papers in this Special Feature offer an important advance in empirical sustainability science research. Like earlier advances in my own field of development economics, however, they suffer some avoidable interpretive and ethical errors. This essay celebrates the powerful potential of design-based sustainability science studies, much of it admirably reflected in this set of papers, while simultaneously flagging opportunities to improve future work in this tradition.

causal inference | policy | randomized controlled trials | sustainable development

An exciting feature of sustainability science is its focus on problems that transcend disciplines, compelling scholars to integrate natural and social sciences. However, the study of closely coupled human and natural systems poses a formidable empirical challenge. We must disentangle multiple plausible causal mechanisms behind the natural and social phenomena we observe and separate true causal effects from spurious correlations generated by unobserved, confounding variable(s). For example, when we observe deforestation coincident with increased diffusion of inorganic fertilizer and improved seeds among smallholder farmers, it could be that deforestation was caused by farmers clearing land for crop agriculture made more profitable by improved agricultural inputs. It is also plausible that both phenomena could be driven by changing market, soil, or weather conditions. Increased uptake of modern agricultural inputs might even be attenuating what would otherwise be even more rapid deforestation driven by rising human population density. Designing appropriate policy to advance the Sustainable Development Goals requires sorting among those candidate mechanisms. We must offer statistically sound answers to policy-oriented questions concerning whether technological, sociocultural, policy, market, institutional, and/or other interventions work, why or why not, and under what conditions.

Generating credible inferences to answer those key policy questions is difficult in complex systems. The sustainability scientist in the field cannot replicate laboratory conditions where she can both directly manufacture variation in the intervention ("treatment") variable of interest and reliably limit confounding variation in outcomes arising due to unobserved variation in other features of the human or natural environment. Closely coupled systems routinely contaminate observational (i.e., nonexperimental) data with statistical endogeneity, which occurs when explanatory variables covary with both the dependent variable and some omitted variable or with the dependent variable due to reverse causality (simultaneity). Returning to the example of deforestation amid smallholder agriculture, both farming and forest clearing behavior might vary in response to some other change—for example, opening a road or a nearby factory—or forest loss might induce change in farmers' agricultural input choices.

Endogeneity biases inferences about policy-relevant parameters. The severity of that bias depends, in large measure, on the researcher's ability to control for prospective confounders, and on the veracity of the model fit to the data. Rigorous deductive or inductive models that explain how and why variation in one human or natural feature changes outcomes take us quite far in

[a]Charles H. Dyson School of Applied Economics & Management, Cornell University, Ithaca, NY 14853

Author contributions: C.B.B. designed research, performed research, and wrote the paper.
The author declares no competing interest.
This article is a PNAS Direct Submission. P.J.F. is a guest editor invited by the Editorial Board.

[1]Email: cbb2@cornell.edu.

Published July 12, 2021.

understanding complex systems, as manifest in pathbreaking sustainability science contributions by giants like Ken Arrow, Bill Clark, Partha Dasgupta, Pam Matson, and Lin Ostrom, among others. But there remains the haunting suspicion that subtle-but-strong endogeneity might induce mistaken inferences.

Sustainability science is not the only problem-oriented field that has faced the challenge of causal inference in the presence of endogeneity. For example, the agricultural and health sciences have long had to identify which new technologies—fertilizers, pharmaceuticals, seeds, vaccines, etc.—generate demonstrable improvements under real-world conditions. Starting more than a century ago with William Gossett's agronomic trials in Ireland, and then Ronald Fisher's trials in England, those problem-oriented research communities gradually embraced design-based research (DBR) methods. The best-known DBR method is the randomized controlled trial (RCT), in which the treatment variable is directly manipulated by the researcher. DBR focuses less on constructing models of complex systems and more on constructing study designs that can isolate and quantify the causal effect of one variable on another. In other words, DBR methods focus on eliminating rival explanations for the patterns observed in data so that one can more confidently interpret about interpreting observed correlations as reflecting causal relationships. Causal interpretations are possible even without a well-developed, model-based theory of why a causal effect may exist. For example, aspirin and urea were found efficacious via DBR before scientists could pin down the mechanisms that caused the observed effects.

DBR contrasts with model-based research (MBR), which focuses less on credible causal inference and more on constructing coherent, compelling models of key mechanisms that regulate the system under study; that is, MBR aims to explain the data-generating process of an outcome variable like deforestation. To oversimplify only slightly, DBR aims to estimate "the effects of a cause," whereas MBR aims to identify the mechanisms that are "causes of an effect."

Sustainability science needs both DBR and MBR. When DBR's demanding underlying assumptions are satisfied (briefly described below), DBR can generate credible statistical inferences about whether the designed variation caused change in the outcome(s) of interest. When MBR's assumptions are satisfied (often involving a combination of assumptions about stochasticity, functional forms, and model constraints), MBR can explain why a causal effect arose and what would have happened had conditions been different.

Over the past quarter-century or so, the international development community and many public and private organizations have adopted DBR, especially the RCT, as a powerful approach for evaluating the impact of specific policy, practice, or product interventions. It is no surprise that sustainability science is starting down the same path.

By learning from both the successes and errors of prior efforts to incorporate DBR, sustainability science can perhaps accelerate useful discovery. The papers assembled in this Special Feature represent an important advance in empirical sustainability science research (1–8). They also, however, repeat some avoidable errors made by prior travelers down this path. This essay attempts both to articulate the powerful potential of DBR in sustainability science, much of it reflected in the papers of this Special Feature, and to flag opportunities to improve future work in this tradition.

Design-Based Empirical Research

This Special Feature reports findings from a set of studies about a specific institutional mechanism—community monitoring—hypothesized to improve common pool resources (CPR) management. As a collection, these papers offer a terrific lens for understanding the possibilities and limits of DBR in sustainability science. In this section, I focus on the value of DBR as a complement to MBR in sustainability science. The next section highlights key insights from the collection in this Special Feature. I do not review the individual papers, already ably summarized in the metaanalysis paper (8). The subsequent two sections explain shortcomings common to this collection of papers, and to much of the earlier development economics RCT literature, so as to help inform emergent DBR in sustainability science.

The power of RCTs arises from the researcher purposefully manufacturing variation in a key explanatory variable. By design, we know that this variation is caused by observed forces external to the system under study. Variation in the explanatory variable of interest is therefore exogenous and not subject to statistical endogeneity concerns that pervade observational data.

This innovation matters in sustainability science because exogeneity is especially valuable in the presence of feedback among subsystems. Feedback poses inferential challenges by increasing the risk of statistical endogeneity. Farmers' responses to changing tree cover and its impacts on soils, pollinators, microclimate, etc. is one example of such feedback. The RCT allows researchers to manage this feedback, obviating endogeneity concerns through a known research design that generates exogenous variation in the key explanatory variable.

DBR more broadly relies on this same exogeneity assumption; the RCT is not the only DBR method. DBR requires neither direct experimenter manipulation of the explanatory variable nor randomization of treatment. Instead, a researcher can exploit other forces (e.g., rainfall, an unexpected policy change) that manufacture known, exogenous variation in a key variable. DBR only requires that the researcher know, and be able to credibly control for, the forces driving variation in the explanatory variables of interest and be able to control for systematic differences in prospective confounders—that is, achieve statistical "balance"—across subsamples. With that knowledge and control, the researcher can develop a credible counterfactual to true observations from which she can credibly infer the causal effects of variation in the plausibly exogenous explanatory variable(s).

Randomized researcher-directed interventions, like the ones used in the Special Feature studies, are a simple way to know the source of variation in the treatment variable and achieve statistical balance. But researchers can also control for endogeneity concerns by using rule-based designs (e.g., observable, exogenous, enforced eligibility criteria for a new program) or natural criteria (e.g., a river or temperature) that affect variation in a treatment variable. For example, a researcher can work with a nongovernmental organization (NGO) to identify the precise criteria that the NGO employs to recruit participants into a program, and then control for the resulting, nonrandom "selection effects" using observations of variables that fully describe those participation criteria (9). This option matters because some operational agencies reasonably object to randomization in program implementation, for ethical or political or targeting reasons (8). With careful attention to how an intervention is designed and implemented, DBR remains feasible beyond the restricted range of partners willing to turn over participant recruitment or intervention design to a research team running an RCT.

Despite the potential contributions of DBR, MBR will always be an essential complement (10). In many important contexts, experimentation or quasi-experimental methods are infeasible or unethical (e.g., understanding the likely impacts of sea level rise or

the introduction of a novel virus or an invasive species). Even where DBR is feasible and ethical, all causal inference fundamentally depends on theory and associated assumptions, whether implicit or explicit, necessarily imposed to make the empirical research tractable (11–16), as I explain in *DBR Interpretation and the Need for Nuanced and Humble Inference*. MBR methods are also essential to understanding why causal effects arise in DBR studies.

DBR Contributions in This Special Feature

The best feature of this set of papers is their metastructure. The researchers undertook a courageous, coordinated effort at design-based replication of tests of a pair of key hypotheses. These hypotheses follow directly from well-developed theory and a wealth of prior observational studies about sustainable and equitable governance of CPRs. Strong, prior MBR laid a solid foundation to hypothesize both that community monitoring is a key feature of effective CPR management and that imposing institutions from outside might prove ineffective (17–29). The team designed a set of replicated interventions wherein outsiders tried to encourage community monitoring of CPRs rather than waiting for select communities to pursue community monitoring endogenously. The core design involved harmonized interventions across 747 communities in six different sites—in Brazil, China, Costa Rica, Liberia, Peru, and Uganda—spanning not only four continents but also distinct CPR challenges: stemming deforestation, groundwater overuse, or surface water pollution.

The interventions were reasonably standardized across sites. Each convened a community workshop to explain the intervention, the CPR management challenge, the state of the CPR of local interest, and monitoring methods. Then a group of individual monitors satisfying prespecified criteria were selected, trained, and, in most cases, provided with some new technology: a water level sensor, a smartphone app that reports remotely sensed measures of deforestation, etc. Monitors were compensated to report the CPR state regularly over the course of at least one year, and the reported information was shared with the community and with a relevant (typically government) CPR oversight body. The objective was to see whether outside intervention could induce increased community CPR monitoring and measurable improvements in resource state. The variation in contexts, CPRs, and the specific interventions is a feature, not a bug. It ensures the metastructure offers a reasonable test of the core hypotheses: Can outsiders effectively induce increased community monitoring, and do such interventions improve the observed state of CPRs and the well-being of households in the treated communities?

In my own field of development economics, this work most reminds me of the seminal six-country, coordinated RCT-based study of so-called "graduation" programs (30). Those programs involved multifaceted interventions intended to stimulate self-reinforcing economic advance among the ultrapoor. The programs generated important research discoveries and have proved hugely influential among donor agencies, governments, and NGOs in the Global South.

The studies in this Special Feature nicely demonstrate three key advantages that DBR offers sustainability science. First, the entire design builds on prior theory and observational empirical evidence, complementing rather than competing with MBR. The authors motivate their DBR by observing that prior MBR findings may be vulnerable to endogeneity concerns and that previous, single-shot, DBR findings may not prove externally valid. This set of studies nicely models sequential learning, in which we learn incrementally from each new data point, some model based, others design based, even if they do not each have equal likelihood of reflecting the truth unbiasedly.

By building on MBR studies, the DBR in the Special Feature is an advance on early DBR in development economics. One of the greatest missed opportunities of development economics' RCT revolution was that many so-called "randomistas" ignored or dismissed findings from prior MBR studies simply because they were not design based. In reading such published RCT-based papers, one sometimes had the sense that the researcher's objective was less to advance collective, cumulative understanding around a real-world problem than to assert individual or small-group intellectual dominance. DBR-MBR competition is counterproductive. Whichever precedes the other—DBR before MBR in the case of early aspirin or fertilizer research, MBR before DBR in modern sustainability science—they complement one another to accelerate understanding. The team of scholars in this Special Feature charted an admirable path different from that in development economics. I pray the sustainability science community will follow their stellar example.

Second, discerning readers should not accept a single set of results as the final word on any empirical question. Lasting scientific discovery comes through replication, from uncovering empirical regularities through repeated study. This research team built in harmonized replication from the outset. Bravo! The scale of the replication increases the likelihood that the researchers uncovered some generalizable patterns. Notwithstanding my cautions about their contestable interpretation of some of their own findings—which I discuss in the next section—their bottom-line results are convincing: Outsiders can stimulate increased community monitoring, and, on average, such interventions measurably improve the state of the CPR. These findings validate prior theory and observational evidence (i.e., MBR) hypothesizing that community monitoring will improve the state of CPR, while MBR concerns that external agents cannot intervene to help boost CPR management and improve outcomes appear incorrect. Those are important lessons well learned.

Third, because research designs based on direct intervention—rather than on natural experiments—are complicated, such scholarship almost always requires collaboration by teams that encompass diverse disciplinary skills and actively engage with subject communities. Researchers cannot help but learn from others on the team—especially in multidisciplinary projects like this one—and from resource community members, especially when compared to distant, disciplinary use of secondary data.

The interdisciplinary teams that carried out these studies exemplify this model. The integration of interdisciplinary and qualitative learning to help inform the design, implementation, and interpretation of a study's quantitative design findings adds considerable value, fosters faster midstudy correctives, and builds in a natural audience for research findings. One of my field's preeminent design-based researchers, Nobel laureate Michael Kremer, emphasizes that "in addition to isolating causal impact, field experiments have four other key features": (ref. 31, p. 1975), They 1) provide a richer sense of context, 2) address very specific, practical problems, 3) foster collaboration, and 4) promote iterative study over single-shot investigation. This set of studies exemplifies these important virtues.

The DBR nicely modeled in this Special Feature offers a step forward for sustainability science. But it is not a step without some stumbles. As other sustainability science scholars embark down this path, they can learn not only from these many admirable features but also from the stumbles to which we now turn.

DBR Interpretation and the Need for Nuanced and Humble Inference

Careful researchers guard against overstepping the support of their data and methods and recognize that all statistical inference relies on untestable assumptions, DBR no less than MBR (11, 14). I raise this caution because, in development economics, the mechanical task of randomization induced much overconfidence in inferences based on RCT results. Avoidable errors of interpretation are perhaps inevitable early in the practice of new methods. After years of thoughtful critique by scholars like Michael Carter, Nancy Cartwright, Angus Deaton, Lant Pritchett, Martin Ravallion, and others, development economists have gradually embraced more nuanced inference, recognizing and acknowledging RCTs' limits even while celebrating the considerable value DBR adds. Hopefully, sustainability science can learn at a faster rate by absorbing some lessons learned in other fields.

In what follows, I focus on RCTs as a specific DBR method, the one employed by the studies in this Special Feature. The twofold purpose of an RCT is to 1) manufacture exogenous variation through random assignment of some intervention to 2) a treatment subsample that is intended to be otherwise identical to the control subsample. Accurate interpretation of the subsequent findings requires that researcher and readers ask themselves, "What is under the experimenter's control, and what is not?"

Implementing a watertight design that manufactures exogenous variation in the otherwise typically endogenous explanatory variable of interest—like community monitoring—is a difficult task. Indeed, for reasons I explain below, manufacturing exogenous variation is especially challenging for information-based interventions of the sort that the studies in this Special Feature employ.

It is likewise difficult to ensure that treatment and control groups are otherwise identical. As explained below, true balance between control and treatment is especially unlikely in natural systems subject to frequent, major political, weather, and other shocks, the sorts of places where much sustainability science field research focuses.

Therefore, although full researcher control over both intervention and all prospective confounders is desirable and feasible, it cannot be taken for granted. Having myself fielded several DBR studies—some successful, others less so—my experience is that adequate control is more difficult and more rare than some RCT enthusiasts admit. Hence the need for caution and nuance in interpretation and for humility in acknowledging that DBR can fall prey to the same endogeneity problems that so often bedevil observational studies.

When researchers and readers lose track of what was and was not under the experimenter's control, they often make statistically indefensible, logical leaps. In seeking an elusive gold standard, the superficial glitter of randomization prompts them to grasp something more like fool's gold: empirical results that suffer biases similar to those in observational data. A raft of papers enumerate these issues in detail, going into greater depth and discussing problems, like Hawthorne or John Henry effects, that I ignore here due to space constraints (11–16, 32–37). In the interests of brevity, I emphasize just four points that seem especially salient to sustainability science and to the studies in this Special Feature.

The first concern is a well-known caution that bears repeating: Randomization's appeal is its asymptotic balance property. As a sample randomly assigned to treatment or control subsamples grows infinitely large, control and treatment groups converge toward mean balance on all potentially confounding attributes. But balance almost surely does not hold in any finite sample. Hence the good practice followed by the Special Feature studies of inclusion of baseline controls, often including lagged values of the dependent variable, to move the analysis closer to that elusive, asymptotic standard and thereby improve inference (38, 39). But the experimenter can only check and control for the prospective confounders that she measures. Statistical tests of differences between control and treatment groups are valid only conditional on the veracity of the maintained, untestable hypothesis that the randomized groups are balanced on all confounding unobservables as well. The plausibility of that assumption increases with sample size and the scope of plausible confounding variables the study measures. Small RCTs with limited measured covariates are especially vulnerable to imbalance on confounding unobservables.

The well-known problems of balance on observables between control and treatment groups at baseline—that is, ex ante to intervention—have an analog, less-recognized problem of ex post imbalance in exposure to stochastic events. Development economists have recognized that covariate random shocks like tropical storms, heat waves, or localized droughts can compromise the external validity of even internally valid RCTs (39). But the problem runs deeper because shocks are unlikely to hit each arm of an experiment identically even if they face similar ex ante risk of shock. What initially appears like a well-implemented, fully-balanced-at-baseline RCT gets compromised if, for example, unobserved hydrological features of sites leave them differentially vulnerable to surface water contamination from sources unrelated to the experiment (e.g., fires or industrial accidents upstream) or if some sites' topography or proximity to the coast leaves them more vulnerable to tropical cyclone damage to forests. Put differently, nature often reintroduces, ex post of assignment, the imbalance that randomization tried to remove at baseline. Oddly, the "radical skepticism of observational research"—the instinct to believe that observational data are plagued by endogeneity problems even when the skeptic cannot pinpoint the mechanism giving rise to those problems (40)—is not matched by comparable skepticism of the absence of unbalanced stochasticity in an increasingly shock-prone world. Like others, I favor uniform skepticism and steady updating of prior beliefs based on theory and evidence of all sorts, recognizing that study findings are not equally credible, but they are all vulnerable, even those based on DBR (11–16, 32–37, 40). Balance on unobserved confounders is essential in design-based studies, but it is rarely fully under researcher control, especially because imbalance can arise after assignment.

My second caution relates to essential or structural heterogeneity (11–16, 32–37). Intent-to-treat (ITT) effect estimates are unbiased representations only of differences in subsample mean effects of the intervention administered by the experimenter, that is, the intended treatment. But an unbiased ITT estimate is a sample-specific, data-weighted average that often masks variability conditional on a range of observable and unobservable attributes. Consider an RCT that randomly assigns farmers to receive improved seed and fertilizer, then tests for differences in forest clearing. The ITT estimate necessarily blends the effects among farmers living on treeless landscapes with those neighboring forests. The effects of treatment almost surely differ among those two structurally different subgroups. If agricultural extension agents or forest conservation authorities would deal with the two types of farmers differently, then the ITT is not very useful, even if it offers an unbiased estimate of the population-scale effect. Theory and prior evidence—that is, MBR—can be especially useful in helping isolate attributes—like proximity to forest—to use in stratified randomization to accommodate structural heterogeneity. But that requires a

larger sample, and thus greater cost and complexity of implementation, in order to ensure adequate within-stratum statistical power to detect effects of contextually meaningful magnitude. Most studies, including those in this Special Feature, do not include multiple structural strata. But then we struggle to provide convincing ex post tests of heterogeneity using baseline covariates.

The three deforestation studies in this Special Feature nicely illustrate the problem of structural heterogeneity. The Peru study finds large average, albeit imprecisely estimated, reductions in deforestation induced by the intervention, with considerable within-sample heterogeneity and no significant spatial spillovers (7). By contrast, the Liberia and Uganda studies (3, 5) find no difference in forest loss. Indeed, in Uganda, deforestation increased in unmonitored areas of treatment communities, more than offsetting the modest reduction in forest loss in monitored areas. The Uganda findings demonstrate the real risk of spatial displacement: Without discouraging forest use generally, spatially restricted monitoring (or sanctioning) may merely reallocate harmful activities across space. Meanwhile, the Liberia results underscore that insufficient CPR monitoring might not be the causal factor behind deforestation in all contexts. These are interesting and important results.

Metaanalysis can only partly resolve the heterogeneity challenge. The summary paper of this Special Feature claims that, by pooling data across sites to boost statistical power, metaanalysis finds that the interventions, on average, reduced CPR use, as measured in standardized effect sizes (8). Perhaps. Without delving into technical details, while metaanalysis is a valuable tool, it is also complex, yielding results that are often quite sensitive to analysts' technical choices, many of which necessarily require untestable assumptions (41). The central tendency reported in the summary paper seems supported at least as well by the more qualitative evidence reported in individual papers as by the statistical metaanalysis. But surely the headline result is the heterogeneous effects of these interventions on CPR state, not a (contestable) finding of a mean metaeffect in a metasample that is exceptionally heterogeneous.

The concern for sustainability scientists is that the more structurally heterogeneous the sample, the less likely that reported mean effects reflect conditions in any one location. That fundamentally limits how much we can learn from DBR alone, even from impressive, elaborate, multisite efforts.

Structural heterogeneity bedevils MBR too, of course. The solution is to blend the two. DBR research that appropriately integrates MBR can sometimes solve the problem with strata-specific effect estimates. But, because DBR in no way obviates the challenge posed by structural heterogeneity, DBR alone is hardly more informative than MBR on its own. DBR does not automatically overcome the inferential problems familiar in observational data, even if hard-core randomistas often ignore the problem.

The third issue concerns how estimated effects may change over time. Here too, DBR is as vulnerable to the problem as MBR. Most interventions are brief, at most lasting a year or two, as in the Special Feature papers. These teach us something, but require cautious interpretation, in two distinct ways.

One must consider carefully not only the intervention duration—is it one-off or sustained, and, if sustained, for how long?—but, perhaps even more, the timescale necessary to discern whether the intervention induced a measurable disruption. In sustainability science, human and natural outcomes of interest commonly change at different time steps. Some slower-changing variables may only manifest change over extended periods. For example, starting in 2009, my colleagues and I undertook RCT-based evaluation of livestock insurance interventions' near-term impacts on household and individual well-being and behaviors in the rangelands of northern Kenya and southern Ethiopia (42, 43). We have only recently begun to try to evaluate the insurance's impacts on rangeland ecosystems, because short-run weather fluctuations and slow change in land cover suggested that ecosystem impacts would likely become detectable on a much slower time scale than impacts on human well-being (44). In complex, closely coupled systems, the slower-changing state variables are sometimes the most interesting and important ones but are easily missed in studies limited by short funding cycles. Sustainability scientists must be especially attentive to this temporal mismatch problem.

Moreover, one needs to think carefully about whether any estimated near-term impact likely represents a nonstationary (i.e., permanent) effect or whether it was merely a transitory, disequilibrium disruption wherein the underlying system—and its constituent subjects—will likely return to prior state in due time. The evaluations of the Youth Opportunities Program in Uganda offer a cautionary tale. A careful RCT-based evaluation found that this cash grants and skills training program achieved excellent results in boosting participants' skilled employment and income two and four years after intervention (45). Follow-up work by the same authors, however, confirmed theory-based concerns that these effects might not last; the gains had largely vanished after nine years as the general equilibrium effects of the program ultimately dwarfed the direct, partial equilibrium effects (46). In some systems, the speed with which the balancing feedback within the system returns it to initial state may extend inconveniently past the end of a research project's funding cycle. One can only interpret the findings as a characterization of time-bound effects, which may not last.

These time scale issues create an opportunity for sustainability scientists, who tend to think hard about state and transition models, and whether or not shocks are likely to transform rather than merely transitorily perturb a system (47–51). Scant DBR in the agricultural or health sciences, or in development economics, studies the path dynamics of effects induced by a designed intervention. Sustainability science may have a comparative advantage in developing DBR around system dynamics, about how to identify whether an intervention persistently transformed the underlying system state or merely perturbed it temporarily. But keep in mind that most researchers cannot control the period of observation, or the pace of system adaptation. Moreover, the longer the time period of study, the greater the likelihood that the DBR suffers from the second problem I discussed: confounding imbalance caused by shocks ex post of assignment to control or treatment.

Fourth and finally, DBR runs a risk of heterogeneous treatments, a risk that is especially salient to sustainability science (6, 16, 33, 52). Note that heterogeneous treatments are not the same as (structurally) heterogeneous effects of a treatment, discussed previously. Rather, the issue is as follows. In a pure RCT, all treatment group participants comply with their assignment. For example, those given a placebo shot do not get the true vaccine, and vice versa. The ITT then offers a direct estimate of the average treatment effect (ATE) of the intervention, that is, the impact of treatment on the treated population.

Often, however, the researcher has no control over the thing one wants to change, for example, community monitoring in these Special Feature papers. The feasible intervention is, instead, encouragement intended to induce the intended behavior. The Special Feature studies' randomized encouragements to community monitoring—technological aids, training for monitors, community

discussion about the importance of CPR and the risks of overuse, etc.—appropriately reflect the feasible tools a policy maker might use to induce community monitoring. The ITT provides a useful, unbiased estimate of the causal effects of these fully controllable interventions. That is good DBR.

But, often, we also want to know the impact of the endogenous explanatory variable—that is, community monitoring—not just of the intervention instrument—for example, community discussion. Here things get trickier because encouragement can and does fail; and how it fails matters for credible inference.

If subjects either comply and do what they are encouraged to do—monitor the CPR—or ignore the encouragement and do nothing, then the binary endogenous response of compliance or noncompliance causes attenuation bias in the ITT estimate relative to the true ATE (i.e., it is biased toward finding no effect). A careful researcher then estimates the local ATE (LATE) by correcting for the probability of compliance. If the randomized encouragement does not have any pathway to CPR impact other than through the behavioral change one intended to induce, then statistical correction for the probability of noncompliance yields an unbiased LATE estimate. The Special Feature studies do this. Problem seemingly solved.

A problem arises, however, if subjects do not ignore the encouragement and are indeed induced to change their behaviors, just not all in the intended way. Now there are more than two options: compliance or noncompliance. There suddenly exist multiple candidate responses to the encouragement design. The compliance problem morphs from binary to multinomial choice. This problem originates in human subjects' agency, and thus is unavoidable. In principle, one could correct for the multinomial choice to retrieve an unbiased LATE estimate. But researchers rarely notice, much less correct for, unintended, multinomial behavioral responses in response to encouragement designs. The Special Feature papers thoughtfully laid out theories of change and tried to monitor each step along that pathway. But they understandably do not report searching for unintended behavioral responses, nor make corrections for unmonitored, unintended actions.

Subjects' unmonitored, unintended behavioral response can confound the LATE estimate. Let me explain with a teaching metaphor. As any experienced lecturer well knows, what different students learn from a single lecture to which they are each identically exposed varies dramatically among individuals, as will the behavioral response that lecture induces in each student. Information-based encouragement treatments are like students' interpretations of and responses to a lecture. The treatment as given by the experimenter is as homogeneous as the single lecture a teacher gave to a set of students. Thus, the ITT estimate of the treatment's impacts retains the same statistical properties that RCT champions celebrate. And that is a policy-relevant parameter because it relates to the thing an external agent—a teacher, an environmental NGO, a government agency—can control.

But the treatment as received by human subjects is almost surely heterogeneous across subjects, perhaps especially so for information treatments. Imagine natural science students experimentally assigned to my economics course. Those in the control condition all complete college without taking an economics course. Those in the treatment group are all directed to take my course, with the intention that they learn economics, and we can evaluate the impact of economics training on natural science students. But one treatment group student concludes, after hearing just my first lecture, that economics—indeed, college education as a whole—is silly. She drops out of college before completing the course, starts a business, and becomes fabulously wealthy. Another treated student flunks my class, and also fails to learn economics. Like the other treated students, who all learned economics as intended, the student who flunked enjoys no resulting gain in wealth. We have two different types of unintended noncompliers in the treatment condition, students who (for two different reasons) did not learn economics despite receiving the treatment.

The ITT estimate from that RCT would correctly find that attending my first lecture caused an average increase in wealth, thanks entirely to the one student who did not comply, and in an unintended and unexpected way. The point of the metaphor is not the risk of outliers in small samples. Rather it is a caution that, lacking a model-based mechanism, one cannot infer from the RCT why attending my course increased students' wealth. The ITT estimate only tells us whether attending my course led to increased wealth, not why, unless one has a credible LATE estimation strategy.

The problem is that few researchers or other readers of scientific studies are satisfied with an ITT result without an explanation of the accompanying mechanism. Inquisitive minds, like nature, abhor a vacuum and typically cannot resist filling in the "why" blank left by the ITT estimate. Many readers would inevitably misinterpret the summary ITT estimate as a finding that learning college economics—the intended endogenous behavioral change—makes the average student wealthier. I refer to this as the "leap-of-faith" estimate because it burdens the unbiased ITT estimate with an added, untested belief that the encouragement treatment induced the intended behavioral response. But the real mechanism leading from treatment to impact is that my economics course induces unintended and undesired behavioral responses that, perhaps counterintuitively, generate the observed impact.

This silly example illustrates how subjects' agency necessitates care in interpreting findings from an RCT that aims to induce behavioral change, just as it does in MBR inference, because human agency is the handmaiden of statistical endogeneity. Designs based around induced behavioral change can only reliably estimate the impact of the intervention as administered (the ITT effect). One cannot assume a mechanism that drives that result. Yet people routinely assume the intended mechanism of the encouragement design holds, and thus subtly transform the ITT estimate into a much less credible leap-of-faith estimate. This is, unfortunately, somewhat true of these Special Feature papers.

We know, from the Special Feature studies, that the harmonized treatment induces a range of indicators of increased community monitoring in treatment villages: more reporting, more villagers indicating they were aware of reports, etc. These results suggest that the intervention induced increased community monitoring. We also know that there seems an average reduction in CPR use. The authors acknowledge that individuals and communities endogenously choose their responses: change their beliefs about the state of the resource, the value of monitoring, how best to enforce use restrictions, etc. The authors admirably outline theories of change and tried to collect data along that whole causal path. That is far more than most development economics RCTs do.

But it is exceedingly difficult to anticipate, measure, and control for all relevant induced behavioral changes. Thus the authors impose on their impact estimates the plausible but untested prior belief that the mechanism from encouragement to CPR impact runs through increased community monitoring. That is, they turn the unbiased ITT into a less credible leap-of-faith estimate based on what is, in essence, observational evidence of the mechanism(s) of impact. To be clear, I'm prepared to believe them for the same

reasons I find many careful MBR studies' empirical findings plausible. I subscribe to the late Ed Leamer's famous caution that "[o]ne should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data" (ref. 11, p. 31). Again, all statistical inference relies on untestable assumptions, even RCTs.

In the agricultural and health sciences, leaps of faith are less commonly necessary. A pharmaceutical or vaccine or seed or fertilizer blend is identical (excepting manufacturing imperfections) as both given to and received by all treated subjects. But, to paraphrase an old adage, there's many a slip twixt giving and receiving information. Even when the connection seems natural between encouragement treatment and behavioral mechanism, subtle confounders can disrupt the unobserved conversion from the treatment as administered to that received. Those confounders are often endogenous to a host of unobservable attributes of the treated unit. This implies nonclassical measurement error in the exogenous treatment variable. Such nonclassical measurement errors are likely correlated with errors in the measured outcome, leading to bias of unknown sign, and where correction of one source of error can aggravate rather than reduce bias (53). Unobservably heterogeneous behavioral response to encouragement thus reintroduces statistical endogeneity that the randomized treatment set out to remove. DBR is still useful in these settings but requires nuance and caution in interpretation.

Sustainability scientists need to remain alert to the lure of the leap-of-faith estimate arising from the researcher's difficulty anticipating all possible endogenous behavioral responses to treatment. One must be careful not to substitute the intent of the treatment—for example, to induce learning of college economics—for the actual, typically unobservable mechanism that governed the subtle transition from the treatment administered to the information received that actually induces behavioral response. Human research subjects' agency moves behavioral mechanisms beyond the researcher's control. Less control should induce more caution and humility in inference.

One can still make inferences, subject to all the familiar caveats, with respect to the impact of the intervention—ITT effect of provision of the encouragement—on behavioral indicators, the CPR state, and household well-being indicators. But we cannot strictly interpret the ITT as describing the hypothesized behaviorally mediated effect of treatment—that is, the externally induced impact of greater community monitoring—without supplementing with additional information: theory-based assumptions, beliefs based on prior observations, ITT results from related variables, etc. We may want to know the impact of community monitoring, and authors may claim that is what they have estimated, but it is only true subject to the veracity of untested beliefs concerning the underlying impact mechanism(s). Their estimated "impact of community monitoring" is a leap-of-faith estimate. I am nonetheless prepared to believe that the theory-based links in the theory of change the authors map between their interventions and the hypothesized mechanism, plus their ITT results on indicators of community monitoring—reports filed, village awareness, etc.—together reflect an actual mechanism. But I doubt that the interventions, as administered, had identical effects across treated subjects and that the only mechanism to CPR state was through community monitoring. With sufficient theory and carefully measured evidence to boost our confidence in the inference, the leap of faith can be acceptable, just as it can be with inference based on good observational data.

Sustainability science should embrace the potential of design-based methods ably illustrated in this Special Feature while remaining alert to their limitations and to key nuances of statistically defensible interpretations. The statistical rigor of the tool under ideal conditions can prompt overconfidence in users when the realities of the field study compromise the purity of the ideal design. And, in my experience, few research designs emerge unchanged and unblemished from their encounter with real field conditions. ITT effect estimates give us unbiased estimates of differences in only subsample mean effects of the intervention and only over the duration of the study, and only under the maintained hypothesis of both ex ante and ex post balance across subsamples. Even then, they cannot, on their own, tell us much about behavioral mechanisms. LATE effects estimates are unbiased only under the additional strong assumption that the experimental treatment had a direct and uniform effect on subjects' endogenous behaviors and that any estimated impact on the outcome of interest operated solely through that intended encouragement channel. The upshot is the need for greater appreciation of both the natural limits to DBR inference and the opportunities DBR invites to think critically about prospective confounders, to map theories of change, and to be alert to unintended responses. Restrained, nuanced, cautious interpretation might be watchwords as sustainability science enters an exciting new chapter integrating DBR and MBR.

Pay More Attention to Ethical Issues

The international development, and now the sustainability science, communities have wisely embraced DBR the way the agricultural and health sciences have. Regrettably, we have been less quick to commit similarly to scrupulous ethical guidelines for responsible DBR. The horrors of human research subjects' mistreatment in the Tuskegee Syphilis Study, and other biomedical trials before and since, led to the 1978 Belmont Report, which articulated core ethical principles for researchers based in the United States and has influenced human subjects protection protocols worldwide (54).

Human subjects protections rest upon a subtle but profound distinction. In contrast to the researcher who passively observes a system under study—be it an organ in the human body or an ecosystem—an experimental researcher actively, purposefully disrupts that system. Her enhanced agency endows her with greater opportunity to harm study subjects and therefore also invests her with heightened ethical responsibilities for any injury subjects suffer as a result of the intervention (34). Those greater responsibilities manifest in the Belmont Report's three bedrock ethical principles: respect for persons, beneficence, and justice. No informed consent or clever study design absolves activist researchers of their responsibilities to do no harm and to compensate for any injury incurred.

The most basic ethical obligation of all researchers is to do no harm to subjects. Ethical concerns with some RCTs published in high-profile economics journals drew attention but scant response or corrective action (16, 32–34, 55–58). We need more serious self-policing within the academy by institutional review boards, professional associations, and journal editors and reviewers to screen out and sanction ethically dubious behavior. That is especially true when well-meaning, well-resourced outsiders experiment on poor communities less able to absorb, litigate, or even protest the harms done by researchers.

The all-too-real risk of tangible harm done to subjects, whether predictable or truly unforeseeable, also underscores the need for ongoing, near–real-time monitoring and reporting—and quick

correctives—for any adverse impacts caused by the intervention. In the agricultural and health sciences, well-designed, ethically defensible RCTs typically build in safeguards automatically. For example, vaccine trial participants are monitored not only for infection by the disease against which they have potentially just been inoculated but also for any adverse health reactions. And they are assured all necessary care in the event of an adverse response. Similar monitoring and rapid correction for unintended consequences seem uncommon in international development and sustainability science RCTs. We could and should do more to ensure accountability of interventionist researchers and their sponsors.

In the context of the CPR community monitoring RCTs, for example, what if an intended encouragement to monitor water quality had induced, instead, the perverse response that a deranged participant, angered to learn of rampant water overuse and contamination by other community members, intentionally poisoned the community water source? This admittedly far-fetched—but unfortunately not unimaginable—outcome thankfully did not occur. But were the experimenters prepared for such contingencies? The common practice of analyzing data after an endline survey offers no midstudy monitoring and reporting for corrective intervention, as is commonplace in biomedical trials. We need more active discussion and promotion of best practices for monitoring and correction for unintended adverse impacts and of experimenters' responsibilities for their subjects' well-being. In sustainability science, such responsibilities may extend beyond human subjects to other sentient beings and, prospectively, to abiotic conditions material to ecosystem dynamics.

As system thinkers, sustainability scientists will, hopefully, be more alert than development economists and agricultural scientists to the need to monitor for unintended injury. The law of unintended consequences and the close coupling of human and natural systems have the joint effect that interventions intended to advance one goal almost surely lead to an adverse effect somewhere else if one bothers to look; trade-offs abound (59). In the restricted space of social and health sciences, there usually exists a reasonably high likelihood that someone within the human population under study notices and articulates concerns around any adverse human impacts arising, so long as sanction-free reporting mechanisms are implemented as required by most institutional review boards. But recognize that inanimate objects—for example, groundwater chemistry and fault line stress—and many animate ones in the system under study will not complain to the experimenters or their supervisors. Design-based researchers actively manipulating a system must be especially broad and vigilant in monitoring for unintended outcomes.

Because sustainability science always involves communities of subjects and multiple objectives, sustainability scientists will, hopefully, prove more alert than more narrow disciplinary researchers to ethical issues arising from the social choice problem, that is, the challenge of collective decision-making among alternative objectives and outcomes. We know, from Arrow's paradox, that there exists no reasonable algorithmic means of translating individual preferences over multiple options into a collective preference ordering (60). Trade-offs always exist among competing legitimate goals.

Advancing rigorous scientific inference is an indisputably worthy societal goal. But it is almost never the only, or even the dominant, objective within the subject community. We may, therefore, sometimes subordinate rigorous inference to other, higher-priority aims. Consider the case of humanitarian allocations in response to disasters. Do we want to know what works best to save lives and relieve suffering at lowest cost? Of course! But the humanitarian response community strongly favors using the imperfect knowledge they possess about what works best to save lives, relieve suffering, and defend human dignity in the face of anthropogenic or natural disasters without relegating some subjects to a control group designed to be deprived of disaster relief. Rather, researchers rely on natural experiments and careful inference using observational data or contextualized structural models to draw defensible causal inferences as to what works best in humanitarian response (61–63). Power imbalances too often let well-financed foreign researchers with high-level connections run roughshod over the legitimate, competing aspirations of subject communities. One good first step is for interventions to begin with community meetings to fully discuss the study and trade-offs it might entail in treatment communities, and clearly expressed exit options subjects can exercise at any time, as seems to have been true of the research in this Special Feature.

A key objective of randomization is to eliminate selection mechanisms that might confound inference. But whether the randomized allocation of the intervention is more or less ethically defensible than nonrandomized alternative depends on contextual details that determine the appropriate counterfactual allocation mechanism(s) and outcome(s). Randomization might enhance procedural and distributive justice if scarce resources are otherwise allocated based on systems of entrenched power and privilege, for example. Alternatively, randomization might result in an unjust process and outcome when the appropriate counterfactual relies entirely on merit and on heterogeneity in expected returns, such that randomization would grant undeserved standing and opportunity to meritless prospective recipients, violating the Pareto principle.

Imagine, for example, that an RCT seeking to understand whether provision of near–real-time remote sensing products induces people to care more about a CPR and to monitor it more closely. A reasonable randomization design would be random assignment of literate community members into a control group that receives nothing or to a monitor group sent color maps of forest conditions on a weekly basis. But, without stratifying based on subjects' unobserved color blindness and assigning all color-blind individuals to the control group, pure randomization would generate attenuation bias by ignoring that no real-world selection mechanism would ask color-blind persons to monitor complex color maps. This purely hypothetical, but unfortunately realistic, example illustrates that randomization is not always the best research design.

Similarly, if the nonrandom allocation mechanism expressly serves goals of distributive or restorative justice, then abandoning those goals in favor of purportedly purer scientific findings may be ethically problematic, reinforcing systemic inequities within subject communities. Some nonrandom mechanisms are fundamentally unjust; others expressly aim to restore justice. Random mechanisms can, at best, stochastically disrupt fundamentally unjust structures. Is it ethical to substitute the random for the nonrandom-but-just selection method? Unfortunately, RCTs rarely start from an ex ante ethical assessment of the likely nonrandom selection mechanisms in the study context. Like more careful monitoring and correction for adverse outcomes, ex ante qualitative research to establish prevailing real-world allocation mechanisms seems an important discussion as RCTs begin to permeate research on CPRs and other social choice issues in sustainability science.

The final ethical issue I raise concerns inclusive publication. Of the 19 coauthors of the core seven papers in this Special Feature that cover six developing countries, only one scholar has a primary appointment and nationality outside a high-income country—and that's a faculty member at a major Chinese university. This is a talented team, and describing its demographics in no way diminishes the quality of their work. But far more can and must be done to engage locally based scholars in field research projects in the Global South. Moreover, it seems unlikely that these studies could have been completed so successfully without substantive engagement by some local resident(s), the sort that satisfies the widely used Contributor Roles Taxonomy standards for authorship (64, 65). As a community, we have been far too extractive for too long.

Broadening local researcher participation represents both a special challenge and an unusual opportunity for large-scale, multiyear field research projects. It's an opportunity precisely because design-based studies encourage collaboration and local engagement. This emergent community in sustainability science could spur more careful attention not just to causal identification but also to inclusion. If ignored, the rise of DBR in sustainability science could equally aggravate preexisting structural inequalities because of the great expense that large-scale, multisite DBR studies entail. Scholars connected to wealthy institutions in high- and middle-income countries—and to major scientific journals and networks—have enormous absolute and comparative advantage in meeting the minimum capital requirements to undertake such work, and considerable professional power to impose their terms on research collaborators. Too often, those terms prove extractive to local subject and research communities. As a community, we—myself very much included—must up our game and hold ourselves, individually and collectively, to higher standards of inclusion.

Inclusion matters not only for just treatment of contributors from the Global South but also for research quality and impact. Context matters enormously for learning about complex coupled human–natural systems. And those of us at wealthy institutions in the Global North typically understand far less about essential contextual details than do counterparts, or even clever students, more deeply embedded in those systems. We can do better DBR—be more likely to identify prospective confounders arising from unintended behavioral response mechanisms, uncontrolled shocks ex post of subjects' assignment to treatment arms, etc.—by being more inclusive.

One of the most powerful ideas in economics—that there exist gains from trade—requires diversity. No gains from trade exist where all persons and organizations are perfect replicates of each other. If the research community is to reap most of the potential knowledge gains—and deliver to society most of the prospective sustainable development impacts—from DBR in the field, we need to better engage with researchers from and in the Global South.

Two Cheers

We should absolutely cheer the advances represented by this Special Feature. But it falls a bit short of meriting three full-throated cheers. The research team earns one robust cheer for modeling how scholars can develop rigorous, multiscale, DBR strategies that build explicitly on extant theory and prior observational evidence to improve our understanding of key questions in sustainability science. And raise a second cheer for the helpful, convincing empirical evidence they generated on the impacts of outsiders' efforts to introduce new technologies to facilitate monitoring and to encourage communities to boost community monitoring of threatened CPRs. The finding that outside interventions can effectively encourage increased community monitoring and improvements in CPR management is convincing, even if the authors overreach their statistical evidence. But perhaps save the third cheer for studies that more appropriately nuance their interpretations of the results generated by encouragement designs, acknowledge the intrinsic inferential limits of all empirical research, both DBR and MBR, and more explicitly attend to the myriad ethical issues made more prominent by interventionist DBR. I hope that sustainability scientists will study these papers, build on the good examples they set, and firm up their weaknesses so as to effectively employ DBR methods—including but beyond just RCTs—to complement the field's solid MBR foundation.

Data Availability. There are no data underlying this work.

Acknowledgments

I thank Arun Agrawal, Marc Bellemare, Brian Dillon, Paul J. Ferraro, Katie Fiorella, Erin Lentz, Annemie Maertens, Liz Tennant, and an anonymous reviewer for helpful comments on an earlier draft and Kaushik Basu, Michael Carter, Mark Constas, John Hoddinott, Sudha Narayanan, Nishith Prakash, and others for helpful discussions that informed my thinking on these issues. Any errors are my responsibility alone.

1 M. Bernedo Del Carpio, F. Alpizar, P. J. Ferraro, Community-based monitoring to facilitate water management by local institutions in Costa Rica. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015177118 (2021).

2 M. T. Buntaine, B. Zhang, P. Hunnicutt, Citizen monitoring of waterways decreases pollution in China by supporting government action and oversight. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015175118 (2021).

3 D. Christensen, A. C. Hartman, C. Samii, Citizen monitoring promotes informed and inclusive forest governance in Liberia. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015169118 (2021).

4 A. Cooperman, A. R. McLarty, B. Seim, Understanding uptake of community groundwater monitoring in rural Brazil. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015174118 (2021).

5 S. Eisenbarth, L. Graham, A. S. Rigterink, Can community monitoring save the commons? Evidence on forest use and displacement. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015172118 (2021).

6 P. J. Ferraro, A. Agrawal, Synthesizing evidence in sustainability science through harmonized experiments: Community monitoring in common-pool resources. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2106489118 (2021).

7 T. Slough, J. Kopas, J. Urpelainen, Satellite-based deforestation alerts with training and incentives for patrolling facilitate community monitoring in the Peruvian Amazon. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015171118 (2021).

8 T. Slough *et al.*, Adoption of community monitoring improves common pool resource management across contexts. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015367118 (2021).

9 R. Rawlins, S. Pimkina, C. B. Barrett, S. Pedersen, B. Wydick, Got milk? The impact of Heifer International's livestock donation programs in Rwanda on nutritional outcomes. *Food Policy* **44**, 202–213 (2014).

10 A. Agrawal, Common property institutions and sustainable governance of resources. *World Dev.* **29**, 1649–1672 (2001).

11 E. E. Leamer, Let's take the con out of econometrics. *Am. Econ. Rev.* **73**, 31–43 (1983).

12 J. J. Heckman, "Randomization and social policy evaluation" in *Evaluating Welfare and Training Programs*, C. F. Manski, I. Garfinkel, Eds. (Harvard University Press, 1992), pp. 201–230.
13 J. J. Heckman, J. A. Smith, Assessing the case for social experiments. *J. Econ. Perspect.* **9**, 85–110 (1995).
14 M. P. Keane, Structural vs. atheoretic approaches to econometrics. *J. Econom.* **156**, 3–20 (2010).
15 A. Deaton, N. Cartwright, Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* **210**, 2–21 (2018).
16 C. B. Barrett, M. R. Carter, The power and pitfalls of experiments in development economics: Some non-random reflections. *Appl. Econ. Perspect. Policy* **32**, 515–548 (2010).
17 G. Hardin, The tragedy of the commons. The population problem has no technical solution; it requires a fundamental extension in morality. *Science* **162**, 1243–1248 (1968).
18 W. C. Clark, R. E. Munn, *Sustainable Development of the Biosphere* (Cambridge University Press, 1986).
19 E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, 1990).
20 J. P. Platteau, Behind the market stage where real societies exist—Part I: The role of public and private order institutions. *J. Dev. Stud.* **30**, 533–577 (1994).
21 J. P. Platteau, Behind the market stage where real societies exist—Part II: The role of moral norms. *J. Dev. Stud.* **30**, 753–817 (1994).
22 J. M. Baland, J. P. Platteau, *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* (Oxford University Press, 1996).
23 A. Agrawal, C. C. Gibson, Enchantment and disenchantment: The role of community in natural resource conservation. *World Dev.* **27**, 629–649 (1999).
24 P. Dasgupta, *Human Well-Being and the Natural Environment* (Oxford University Press, 2001).
25 A. Agrawal, Sustainable governance of common-pool resources: Context, methods, and politics. *Annu. Rev. Anthropol.* **32**, 243–262 (2003).
26 C. C. Gibson, J. T. Williams, E. Ostrom, Local enforcement and better forests. *World Dev.* **33**, 273–284 (2005).
27 E. Ostrom, M. A. Janssen, J. M. Anderies, Going beyond panaceas. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15176–15178 (2007).
28 A. Chhatre, A. Agrawal, Forest commons and local enforcement. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13286–13291 (2008).
29 O. R. Young *et al.*, Moving beyond panaceas in fisheries governance. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9065–9073 (2018).
30 A. Banerjee *et al.*, Development economics. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science* **348**, 1260799 (2015).
31 M. Kremer, Experimentation, innovation, and economics. *Am. Econ. Rev.* **110**, 1974–1994 (2020).
32 A. Deaton, Instruments, randomization, and learning about development. *J. Econ. Lit.* **48**, 424–455 (2007).
33 C. B. Barrett, M. R. Carter, "Retreat from radical skepticism: Rebalancing theory, observational data and randomization in development economics" in *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, D. Teele, Ed. (Yale University Press, 2014), pp. 58–77.
34 C. B. Barrett, M. R. Carter, Finding our balance? Revisiting the randomization revolution in development economics ten years further on. *World Dev.* **127**, 104789 (2020).
35 A. Deaton, "Randomization in the tropics revisited: A theme and eleven variations" in *Randomized Controlled Trials in the Field of Development: A Critical Perspective*, F. Bedecarrats, I. Guerin, F. Roubaud, Eds. (Oxford University Press, 2020).
36 M. Ravallion, "Should the randomistas (continue to) rule?" in *Randomized Controlled Trials in the Field of Development: A Critical Perspective*, F. Bedecarrats, I. Guerin, F. Roubaud, Eds. (Oxford University Press, 2020).
37 M. Ravallion, Highly prized experiments. *World Dev.* **127**, 104824 (2020).
38 D. McKenzie, Beyond baseline and follow-up: The case for more T in experiments. *J. Dev. Econ.* **99**, 210–221 (2012).
39 M. R. Rosenzweig, C. Udry, External validity in a stochastic world: Evidence from low-income countries. *Rev. Econ. Stud.* **87**, 343–381 (2020).
40 S. C. Stokes, "A defense of observational research" in *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, D. Teele, Ed. (Yale University Press, 2014), pp. 33–57.
41 T. Greco, A. Zangrillo, G. Biondi-Zoccai, G. Landoni, Meta-analysis: Pitfalls and hints. *Heart Lung Vessel.* **5**, 219–225 (2013).
42 N. D. Jensen, C. B. Barrett, A. G. Mude, Cash transfers and index insurance: A comparative impact analysis from northern Kenya. *J. Dev. Econ.* **129**, 14–28 (2017).
43 K. Tafere, C. B. Barrett, E. Lentz, Insuring well-being? Buyer's remorse and peace of mind effects from insurance. *Am. J. Agric. Econ.* **101**, 627–650 (2019).
44 B. T. Bestelmeyer *et al.*, "State and transition models: Theory, applications, and challenges" in *Rangeland Systems: Processes, Management and Challenges*, D. Briske, Ed. (Springer, 2017), pp. 303–345.
45 C. Blattman, N. Fiala, S. Martinez, Generating skilled self-employment in developing countries: Experimental evidence from Uganda. *Q. J. Econ.* **129**, 697–752 (2014).
46 C. Blattman, N. Fiala, S. Martinez, The long-term impacts of grants on poverty: Nine-year evidence from Uganda's Youth Opportunities Program. *Am. Econ. Rev. Insights* **2**, 287–304 (2020).
47 C. S. Holling, Resilience and stability of ecological systems. *Annu. Rev. Ecol. Syst.* **4**, 1–23 (1973).
48 B. Walker, C. S. Holling, S. Carpenter, A. Kinzig, Resilience, adaptability and transformability in social–ecological systems. *Ecol. Soc.* **9**, 5 (2004).
49 B. Walker *et al.*, A handful of heuristics and some propositions for understanding resilience in social-ecological systems. *Ecol. Soc.* **11**, 13 (2006).
50 C. Perrings, Resilience and sustainable development. *Environ. Dev. Econ.* **11**, 417–427 (2006).
51 C. Folke. Resilience (Republished). *Ecol. Soc.* **21**, 44 (2016).
52 C. B. Barrett, A. Islam, A. Malek, D. Pakrashi, U. Ruthbah, Experimental evidence on adoption and impact of the system of rice intensification. *Am. J. Agric. Econ.*, in press.
53 K. A. Abay, G. T. Abate, C. B. Barrett, T. Bernard, Correlated non-classical measurement errors, 'Second best' policy inference, and the inverse size-productivity relationship in agriculture. *J. Dev. Econ.* **139**, 171–184 (2019).
54 United States National Commission for the Protection of Human Subjects of Biomedical, Behavioral Research, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (United States National Commission for the Protection of Human Subjects of Biomedical, Behavioral Research, 1978).
55 S. J. Baele, The ethics of new development economics: Is the experimental approach to development economics morally wrong? *J. Philos. Econ.* **7**, 1–42 (2013).
56 D. L. Teele, Ed., *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences* (Yale University Press, 2014).
57 F. Bédécarrats, I. Guérin, F. Roubaud, Eds., *Randomized Control Trials in the Field of Development: A Critical Perspective* (Oxford University Press, 2020).
58 R. Khera, Some questions of ethics in RCTs. *SSRN* [Preprint] (2021). http://dx.doi.org/10.2139/ssrn.3780908 (Accessed 29 June 2021).
59 M. Herrero *et al.*, Articulating the effect of food systems innovation on the sustainable development goals. *Lancet Planet. Health* **5**, e50–e62 (2021).
60 K. J. Arrow, *Social Choice and Individual Values* (John Wiley, 1951).
61 E. C. Lentz, S. Passarelli, C. B. Barrett, The timeliness and cost-effectiveness of the local and regional procurement of food aid. *World Dev.* **49**, 9–18 (2013).
62 Y. Yang, J. Van den Broeck, L. M. Wein, Ready-to-use food-allocation policy to reduce the effects of childhood undernutrition in developing countries. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 4545–4550 (2013).
63 A. Nikulkov, C. B. Barrett, A. G. Mude, L. M. Wein, Assessing the impact of US food assistance delivery policies on child mortality in northern Kenya. *PLoS One* **11**, e0168432 (2016).
64 A. Brand, L. Allen, M. Altman, M. Hlava, J. Scott, Beyond authorship: Attribution, contribution, collaboration, and credit. *Learn. Publ.* **28**, 151–155 (2015).
65 M. K. McNutt *et al.*, Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2557–2560 (2018).