SPECIAL FEATURE: INTRODUCTION

# Synthesizing evidence in sustainability science through harmonized experiments: Community monitoring in common pool resources

Paul J. Ferraro[a,b,1] and Arun Agrawal[c,1]

Over 30 y ago, Elinor Ostrom published *Governing the Commons* (1), a demonstration by counterexample that the successful management of common pool resources requires neither individual private property rights nor central government control. Her contribution and subsequent research identify more than a dozen institutional features, or "design principles," purported to make successful community-based, common pool resource management more likely (2, 3). Examples of these features include clearly defined group and resource boundaries, graduated sanctions, conflict-resolution mechanisms, enabling policy environments, and accountable monitoring systems. But after decades of theoretical and empirical studies, little is known about whether external interventions can enhance these features where they are weak or absent, or whether enhancing these features individually—rather than collectively—causes resource conditions to improve (4). These questions have long been identified as important for advancing sustainability science and practice (5).

In this Special Feature on "Sustaining the Commons," the authors try to answer these questions by developing and testing mechanism-based theories of institutions in complex socio-ecological systems. The Special Feature makes three contributions to advancing sustainability science. It offers an exemplar of multiteam, harmonized research to accelerate the accumulation of generalizable knowledge in sustainability science. In addition, it sheds light on the causal relationship between community monitoring and common pool resource outcomes, a relationship that has long been of interest in the sustainability science literature (1, 6, 7). Finally, it offers a theoretically informed approach to guide synthesis of research findings from diverse contexts. Through these contributions, the Special Feature aims to stimulate discussion and debate among different communities of scholarship and practice in sustainability science. These communities include researchers and practitioners interested in common pool resource governance, adaptive management, co-management and decentralization, citizen science, and democratic accountability, as well as those interested in applying new approaches for empirical research on causal relationships in complex systems.

The Special Feature comprises six, coordinated country-level studies (8–13) and a metaanalysis of these studies (14). Accompanying these empirical contributions are two perspectives essays authored by scholars outside of the coordinated research project (15, 16). This introductory article first discusses how multisite, harmonized approaches can advance sustainability science. We then turn to the multifaceted, multimechanism construct of "community monitoring" and the role it plays in four different sustainability science literatures on citizen science, adaptive management, common pool resources, and democratic accountability. By highlighting its role in these different literatures and how the empirical studies included in the Special Feature take the insights of these literatures into account, we hope for greater cross-fertilization across these literatures. Finally, we examine challenges to developing generalizable knowledge about the role of institutions in complex social-ecological systems and show how mechanism-based tests of theories can help generate and synthesize empirical evidence in sustainability science.

## Harmonized Studies (Metaketas) in Sustainability Science

Sustainability science promises to address the challenges of sustainable development, which can be defined as "enhancing human well-being to more equitably meet the needs of both current and future generations" (17). But the promise of sustainability science will be hard to deliver without a deeper, causal understanding of how humans and the environment change in mutual interaction. Ideally, we want to know whether (and why) a relationship such as "a change in *A*

[a]Carey Business School, The Johns Hopkins University, Baltimore, MD 21202; [b]Department of Environmental Health and Engineering, a joint department of the Bloomberg School of Public Health and the Whiting School of Engineering, The Johns Hopkins University, Baltimore, MD 21212; and [c]School for Environment and Sustainability, Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI 48109

Author contributions: P.J.F. and A.A. wrote the paper.

The authors declare no competing interest.



[1]To whom correspondence may be addressed. Email: pferraro@jhu.edu or arunagra@umich.edu.

Published July 13, 2021.

leads to a change in *B*" holds in general. If the relationship only holds under some conditions, we want to know under what conditions it holds.

When seeking to shed light on causal relationships using data, sustainability scientists often face tradeoffs between internal validity ("Do my inferences approximate the truth?") and external validity ("Do my inferences generalize to other settings?"). Studies that use the entire Earth as a case study ("a global analysis") offer insights on causal relationships that may generalize across space and, with longitudinal data, time. These global studies also allow scholars to assess how relationships vary across space and time. Yet the myriad assumptions that go into such analyses make it difficult to assess the credibility of the authors' causal inferences (i.e., the study's internal validity). Smaller-scale studies may yield more credible inferences, but the specifics of the study location, measures, and analyses make generalizability and extrapolation of those inferences more challenging.

One path toward resolving these tradeoffs is to aggregate the results from comparable smaller-scale studies. When smaller scale studies sample from a wide range of conditions, aggregating their results can yield a clearer picture of the nature and magnitude of a causal relationship, the generalizability of the relationship and its magnitude across contexts and, in some cases, the factors that lead to heterogeneity in magnitude across contexts. For empirical studies, the most common form of aggregation is a metaanalysis. Yet in sustainability science, like in other scientific fields, there are two obstacles to metaanalyses.

First, the studies that serve as inputs for metaanalyses may be unsuitable. Metaanalyses take studies with high internal validity and enhance their external validity and statistical power. Metaanalyses do not, however, correct for hidden biases in studies with weak internal validity. In the sustainability sciences, hidden biases are widespread and advanced approaches to addressing them remain rare (18). Problems of hidden bias are compounded by widely acknowledged and documented problems associated with "researcher degrees of freedom" (19, 20). Researchers in many disciplines face incentives to generate dramatic findings with few rival explanations. Such incentives lead to a wide range of selective reporting behaviors, including publication biases (not publishing studies that fail to find clear, dramatic results), p-hacking (selectively reporting statistically significant results), and HARKing (changing the hypotheses after the results are known) (21, 22). These behaviors impede clearer understandings of how the world works. Aggregating studies affected by these behaviors through a metaanalysis cannot yield accurate, generalizable knowledge.

Second, even when metaanalyses draw on multiple studies with high internal validity, they generally assume that the included studies focus on the same construct (e.g., community monitoring) and that their outcome measures are comparable after standardization for differences in units (e.g., resource extraction). However, researchers are rarely rewarded for replicating prior studies. Thus, new studies often use designs that differ from prior designs in meaningful ways, including what research question is posed, what constructs are measured, how constructs are measured, and how data are analyzed. These variations make comparing studies—and thus accumulating knowledge—difficult. Solving this problem requires changing the incentives that researchers face by reducing the rewards for novelty and increasing them for harmonization of measures and methods across studies.

Even when scholars can successfully aggregate the results of smaller studies, the individual studies are still important in their own right. Generalizability is enhanced when we understand why a causal relationship was found. Such an understanding is often best provided by individual studies. Moreover, individual studies provide contextual details that are missing from metaanalyses. These details not only help scholars make better sense of a metaanalysis, but they also are critical fodder for theory refinement and hypothesis development.

This Special Feature offers one solution to the tradeoffs between internal and external validity and between a metaanalysis's big picture from a 10,000-m view and a smaller-scale study's detailed picture from a 10-m view. The teams' solution is called a "metaketa," derived from the Basque word for "accumulation." A metaketa, according to The Evidence in Governance and Politics (EGAP) network that developed and championed the idea, is "a collaborative research model aimed at improving the accumulation of knowledge from field experiments on topics where academic researchers and policy practitioners share substantive interests" (defined in the Metaketa Initiative, https://egap.org/our-work-0/the-metaketa-initiative/). Metaketas are similar in some key respects to "coordinated networks" and "distributed experiments" in ecology (e.g., refs. 23 and 24), or multisite randomized controlled trials (RCTs) in health and social policy fields (e.g., ref. 25). Metaketas are also similar in important ways to the "most similar systems" and "most different systems" designs used to explain social phenomena in comparative analysis (26, 27). The metaketa's randomization of a treatment allows the researcher to contrast two similar groups that have different values of a causal variable ("most similar"); its replications across contexts allow the research to contrast diverse groups that have the same value of a causal variable ("most different").

Yet metaketas expand on these existing approaches by grounding themselves in eight principles (see https://egap.org/our-work-0/the-metaketa-initiative/ and ref. 28 for further details), which we distill into four principles:

1) Harmonization: Harmonize research questions and study designs across study sites.
2) Constraints on researcher degrees of freedom: Preregister designs, create preanalysis plans, execute independent, third-party replication of analyses, and archive data and analysis code in public repositories.
3) Synthesis: Execute a preregistered metaanalysis organized around the harmonized research questions and designs.
4) Integrated publication: Present results on a common timeline, and ideally in a common venue, to reduce the risks associated with publishing a second, or subsequent, study on a topic (i.e., to reduce the premium on "novelty" in scientific publications).

Adherence to these principles forced the study teams to harmonize their designs in multiple dimensions prior to implementing their field interventions. First, the teams had to pose the same research questions, guided by a common overarching theory (i.e., every team's theory of change had to be nested within a larger, common theoretical framework). Second, the teams had to design comparable community monitoring interventions and select comparable measures of mechanisms and outcomes, as well as variables hypothesized to moderate the intervention's effects. Each team could pose additional research questions and study additional interventions or variables but had to adopt a core set of harmonized study design attributes.*

*The entire project (six country projects plus metaanalysis) cost USD $1.47 million, of which about $912,000 was for the interventions and the rest for measurement. Based on conversations with EGAP leadership, we estimate that the cost of harmonization represented less than 5% of measurement expenditures.

This harmonization process facilitated the aggregation of the country-level results in a metaanalysis, which synthesized the estimated effect sizes and greatly improved the modest statistical power of the country-level studies. Yet, even in the absence of the metaanalysis, the harmonization process enhanced the external validity of the studies in two ways.

First, because the studies had an overlapping set of research questions, interventions and measures, a reader can more easily compare the results from the individual studies. This benefit, however, comes at a cost: a harmonized design can limit the diversity of interventions that can be implemented across sites. Each team must create an intervention that is recognizable as the "same" across many countries. If the "best" intervention at a site deviates too far from the underlying construct that the metaketa wishes to test, some sites may not get the best intervention. For example, if the best community-monitoring programs require codesign by community members, that requirement may be incompatible with the metaketa's requirement that the monitoring programs be harmonized across sites (codesign of interventions may also be incompatible with creating an intervention that can be easily scaled across many communities).

The second way in which harmonization enhances external validity is through the process by which teams agree on a transparent theory that motivates an intervention. To address prior critiques that experimental designs are often atheoretical and fail to shed light on mechanisms (29, 30), the teams developed a clear theory of change. A theory of change is a hypothesized, mechanism-based causal path from the intervention to the outcomes. To shed light on whether the mechanisms along the hypothesized causal path were operative in the anticipated directions, the teams also engaged in complementary data collection. Ensuring the analysis was well integrated with theory allows readers to be more confident in extrapolating the results to other contexts. Moreover, integration with prespecified theory enhances internal validity. As noted by Barrett and Carter (29), "results are credible not just because they are statistically significant, internally unbiased parameter estimates, but because they conform with a compelling model of behavior."

To enhance the internal validity of the studies, all teams used a mixed-methods approach that relied on randomization of the intervention by the researchers (by "mixed methods," we mean a mixing of quantitative and qualitative data). As with the benefits of harmonization of designs across sites, the benefits of randomization do not come without cost. In a perspective essay in this Special Feature (16), Barrett discusses the strengths and limitations of experimental designs in the context of sustainability science and what we can learn from other fields to improve the application of experimental designs in sustainability science. Moreover, randomized experimental designs are subject to ethical assessments that are often not applied in nonexperimental implementations of similar interventions (31). In another perspective essay in this Special Feature (15), Asiedu et al. discuss ethical aspects of social policy experiments and propose steps that study teams can take to assess the ethical dimensions of their studies prior to implementation and to communicate these dimensions when reporting the results [Barrett (16) also discusses the ethics of social experiments].

The metaketa harmonization process also helped to constrain "researcher degrees of freedom" at the analysis stage, whereby a study's purpose and target parameter can be reinterpreted after observing the initial results. The harmonization process required the teams to develop and preregister their research questions, their analysis plans, and their motivating theory of change prior to the analysis, rather than during or after the analysis. Preregistration required the teams to operationalize theoretical constructs in advance: What is the causal variable (the treatment), what is the target casual effect (the estimand), what are mechanisms that may mediate the causal effect, and what are the moderators that may enhance or diminish the causal effect? The study teams also committed to third-party replication of the results (part of the metaanalysis) and archiving of data and analysis code in a public repository (https://osf.io/). Equally important, the authors agreed to report in their publications any deviations from their preregistered plans.

In some ways, the apparent "noisiness" of the results in the country-level studies (i.e., imprecise or conflicting estimates) may not be simply a function of modest statistical power. It may also be a function of the degree to which the details of the randomized designs and data analyses were prespecified in advance and could not be "massaged" after the data had been evaluated, unless such massaging took place in clearly marked exploratory research sections of the articles (rather than in the introductions, masquerading as confirmatory research). The preregistration ensured that authors could not, for example, see that an outcome variable generated a null result, or yielded a puzzling result that did not conform to the original theory, and then decide, unobserved by readers, that the outcome variable was a "poor measure" and should be dropped from the analysis.

Requiring preregistration of study designs, as well as third-party replication and public posting of all data and code, can thus constrain the questionable research practices that undermine the internal validity of empirical analyses in many fields (32, 33). However, although preanalysis constraints may be a good thing in the long run because they make science more credible, they could, in the short run, further marginalize science in the policy process because they often reduce the certainty about the inferences that can be delivered to policy makers (34).

A final feature of the metaketa in this Special Feature may have further enhanced the internal validity of the designs. The research question was arrived at via a competitive process created by the funder, the United Kingdom's Department for International Development, and managed by EGAP. In this process, the funder identified two broad classes of interventions, information provision and community monitoring, and indicated it was interested in funding experimental evaluations of these interventions in the common pool resource contexts of forests and water. Teams competed in a preliminary proposal phase, after which a subset of teams was selected to move to a harmonization phase around a single intervention. This process led to an interesting outcome: none of the study authors in this special issue are leaders in the four overlapping literatures that have reported on the successes of community monitoring (Table 1). This outcome offers advantages (e.g., new perspectives) and disadvantages (e.g., mischaracterizing the literature), but one important advantage is that none of the study teams had a horse in this race and thus they may have felt less pressure to deliver a particular result. We next turn to the topic of the Special Feature studies: community monitoring.

## Multifaceted, Multimechanism Community Monitoring

All monitoring generates information. This information can improve science, solve coordination problems, generate compliance, strengthen institutions, and support adaptive management (35). Indeed, the large literature on monitoring documents the diversity of uses that information from monitoring enables (36). Monitoring is the foundation for a better understanding of social and environmental systems. Without the information monitoring

**Table 1. Key features of writings on community monitoring**

| Feature | Citizen science | Adaptive management | Common pool resources | Democratic accountability |
|---|---|---|---|---|
| Central puzzle to address via community monitoring | Improve scientific knowledge of resource systems and environment | Improve system performance through better management | Secure and promote user compliance with rules | Hold authorities accountable |
| Role of monitoring | Provide information for better scientific knowledge | Provide information on system outcomes to improve system management | Provide information on rule compliance to better target sanctions and to resolve disagreements | Increase transparency of decision-making |
| Target of monitoring | Species, ecosystems, ecosystem processes | System performance and its relationship to interventions | User behavior, match between user behavior and rules | Authorities, actions of authorities |
| Monitoring mechanisms | Volunteer monitoring networks, campaigns, species counts and lists, mapping games, smartphone apps | User-provided information, specialized monitors, automated sensors | In-person direct and indirect observations, remote sensing data | Citizen scorecards, community meetings, community audits, community-based budgeting |
| Anticipated effects | Changes in scientific knowledge, higher citizen participation, more effective data collection tools | Changes in management interventions and system processes | Increased rule compliance | Greater transparency |
| Ultimate outcomes | Improved science | Improved resource system outcomes | Improved resource system outcomes | Lower corruption |

affords, it would be difficult to assess the speed of climate change, challenging to understand the effects of socio-environmental changes, and impossible to assess the performance of people or programs (37).

In the context of social-ecological systems such as resource commons, community monitoring generates three types of information. It produces information about the state of ecosystems, resources, and species. It produces information about the actions of members of a community or organization, especially whether and how their actions align with explicit or implicit rules governing their behaviors. Finally, it produces information about the actions of elected or appointed authorities and the effects of their actions on people and ecosystems. Individuals, communities, and organizations use these types of information to improve scientific understanding of systems as well as to improve system performance and outcomes via better-designed interventions.

Information from community monitoring is a focus of inquiry in four distinct literatures: citizen science, adaptive management, common pool resource governance, and democratic accountability. As mentioned above, each seeks to achieve information-oriented objectives but to differing degrees and in different ways (Table 1). Their insights also informed the design of the metaketa and the specification of the mechanisms through which the experimental treatment in the studies—community monitoring—could affect resource outcomes.

The citizen-science literature highlights how community monitoring by citizen volunteers can advance scientific goals of data collection, discovery, engagement, and scientific education (38, 39). Scientific monitoring of the status and trends of ecological systems, and the threats to these systems, is costly. Involvement of citizen volunteers reduces costs, but the resulting data may be noisy or biased. A substantial proportion of citizen science research investigates how data collected by volunteer participants can be made more reliable (consistent) and valid (accurate) (40, 41). This literature has expanded considerably in recent years with the emergence of novel partnerships that seek to improve the quality of citizen science by combining large-scale data collection by community members with new sensors, instruments, and analysis techniques (42, 43). Information provided by citizen monitors is improving the quality of science and promising to yield better-designed interventions in fields as diverse as ecosystem services, disease surveillance, sustainable development, climate action, disasters, environmental (air and water) quality, agriculture, urban ecology, and invasive species (44, 45).

In contrast to the emphasis of the citizen-science literature on the use of community monitoring to improve science, the literature on adaptive management emphasizes the role of monitoring in helping elected and appointed authorities make better decisions, thereby improving system outcomes (46, 47). Information about social-ecological system changes, including changes that occur as a result of management interventions, helps resolve uncertainties and change management direction as needed (48). Information is thus of central importance to all adaptive management (49). Although information may be available from diverse sources, community monitors can have an edge over external monitors because they can be less expensive and because they often possess better time- and place-specific information about system changes. In addition, the participation of community stakeholders in monitoring can change their beliefs and resource use behaviors directly (50).

Whereas the adaptive management literature implicitly assumes that decision-makers act in the public's best interests and need better information to do so, two other literatures focus on the role of information under conditions where the private incentives of decision-makers are not aligned with the provision of a public good. The common pool resource governance (or common-property) literature emphasizes how monitoring can align the behaviors of resource users with rules for resource use, while the democratic accountability literature emphasizes how monitoring can change the behaviors of elected or appointed authorities. In small decentralized resource use and management contexts, authorities and users are distinguished mainly by the roles they play with respect to resources: authorities create and enforce rules for the use and governance of the resource system; users rely on the resource system for their needs and are expected to follow collectively agreed-upon rules.

Scholarship on common property has focused on the role of monitoring to address information gaps about user behavior, and

how such information may solve the key problem of user compliance with rules (51). Information helps authorities shape user incentives so that users adjust their resource-related behaviors. In this context, monitoring activities are typically nested in institutional arrangements that include sanctions and adjudication (1). Yet studies, especially those based on laboratory experiments, also suggest that shared information by itself can affect what users do, even without changes in sanctioning and adjudication. The Special Feature studies take this insight a step further in diverse empirical contexts, showing how and to what extent monitoring in common pool resource systems by itself affects outcomes. In addition, the studies elaborate on how monitoring works, a missing piece in writings on common property despite their attention to the importance of community monitoring.

Monitoring in groups—of both users and authorities—raises the complex question of who will monitor the monitor (52). Most groups address this monitoring dilemma through a combination of strategies: role specialization, supervision, and recursivity. With role specialization, a much smaller number of group members—those performing the task of monitoring—need supervision and, in many instances, additional incentives. Elected or appointed authorities often serve as supervisors of monitors. Recursivity occurs when officeholders are accountable to their constituents. In such arrangements, widespread in small-scale common property systems across the low and middle-income world (53, 54), specialized monitors observe and report on the status of the resource and actions of resource users, authorities at different levels supervise monitors, and group members can observe and seek information on the decisions of authorities.

This last form of monitoring—of authorities by group members—is the focus of the literature on democratic accountability (55–57), especially in decentralized contexts. Decentralization of decision-making itself has attracted substantial attention, with some observers calling it "the quiet revolution" (58) and others referring to it as the most important governance trend in the last half century apart from democratic transitions (59). Monitoring for democratic accountability focuses on using information, and transparency through information sharing, to improve governance outcomes (60). Advocates of community monitoring suggest that elected and appointed authorities care about who knows what about their actions and how citizens perceive and judge their actions. Greater transparency and more information about decision-making are thus viewed as critical to hold authorities to account by limiting the use of public goods for private benefits.

Community monitoring approaches in writings on democratic accountability include a variety of mechanisms that may advance transparency of decision-making (61, 62). In studies on the effectiveness of these mechanisms in improving accountability of authorities, the mechanisms may be studied as standalone interventions or as part of a set (63). At the community level, these mechanisms include information meetings, community project monitoring and reporting, participatory budgeting, community scorecards, and grassroots audits (64, 65). Some of these mechanisms may activate greater accountability of authorities through channels other than information acquisition and sharing.

Viewed through the lens of these four distinct literatures on community monitoring (Table 1), the empirical context of the six RCTs in the metaketa most closely resembles the contexts that scholars of common pool resources analyze. But the metaketa's underlying theory of change linking monitoring to resource outcomes draws upon insights from all four community monitoring literatures. The literatures on citizen science and adaptive management both emphasize the role of monitoring in uncovering information to address knowledge gaps of managers and the use of information to improve decision-making. Scholarship on common pool resources highlights the importance of monitoring information and its public availability for changing user behavior, even without externally imposed sanctions. Finally, the democratic accountability literature highlights how information can prompt authorities to make decisions, such as to enforce resource-use rules, for greater public benefit.

## What Can the Special Feature Tell Us about Community Monitoring?

The fundamental challenge confronting the study of institutional interventions like community monitoring is that such interventions are inevitably heterogeneous. In each of the four distinct literatures described in the previous section, community monitoring has different objectives, forms, target outcomes, and hypothesized mechanisms. Compounding these differences are the many potential variations in how community-monitoring programs are operationalized in field settings. Community monitoring may vary across study sites in terms of: 1) the types of information monitors collect; 2) the costs of monitoring; 3) the means of monitoring (how information is collected); 4) the timing, frequency, and sequencing of monitoring activities; 5) how monitoring information is aggregated and analyzed; and 6) who receives the information (Table 2). Other potentially consequential differences in community monitoring interventions may concern the characteristics of monitors, the reliability of monitoring information, or the uses to which the monitoring information is put.

Researchers using observational designs have little control over these variations. But even in harmonized, experimental designs, the implemented intervention may differ across study settings as a result of the variations in the resource system, the socio-economic and political features of communities, and how field teams interpret and implement the harmonized theory of change. The six teams in this Special Feature sought to implement the same community-monitoring intervention, basing their intervention designs on how community monitoring is described in writings on common pool resources and democratic accountability. Nonetheless, as illustrated in Table 2, the interventions vary across many dimensions (for other variations across the studies, and also the key common features, see table 1 in ref. 14).

The challenge for sustainability scientists is to understand the implications of such treatment variation for study designs and analyses (we use the word "treatment" as a synonym for "causal variable," regardless of whether the variable is manipulated in an experiment or not). The challenge for practitioners is to understand the implications of treatment variation for the design and implementation of field interventions. For community-monitoring programs, the observable variations can easily be documented, as in Table 2. Indeed, many metaanalyses aiming to make broad generalizations from quantitative work in a subject area recognize that variations in implemented treatments can explain variations in study findings (so-called "clinical heterogeneity") (66). But to generalize a causal relationship from multiple studies where the implemented treatment is not exactly the same across studies—in this case the effect of community monitoring on common pool resources—more than a taxonomy of treatment variations is necessary. We need to know whether these variations are "consequential" in terms of the relationship between the treatment and the outcomes of interest.

**Table 2. Variations in community monitoring interventions in the Special Feature studies**

| Attributes of intervention | Buntaine et al. (8) | Bernedo del Carpio et al. (9) | Cooperman et al. (11) | Eisenbarth et al. (12) | Slough et al. (13) | Christensen et al. (10) |
|---|---|---|---|---|---|---|
| Resource system | Surface water in China | Groundwater in Costa Rica | Groundwater in Brazil | Forests in Uganda | Forests in Peru | Forests in Liberia |
| Type of information | Odor, color, chemical composition of water | Service quality, water quality, leaks, illegal water/land uses | Well pump electricity use, depth to well water | Tree and branch cutting, grazing, charcoal making, clear-cutting, infrastructure expansion, wildfires | Deforestation incidents and amounts | Scale and types of user activities in forests |
| Cost of monitoring | $2 per session to monitor | $3 per report to monitors | $3 to $5 per month | $2.79/d | $80 per month | $1.75 per month per monitor |
| How information is collected | Volunteers collect data and transmit it | Smartphone monitoring application; WhatsApp chat group | Water committee members collect electricity use and water level data, send WhatsApp reports | In-person patrols along transects | Satellite data, patrols | Remote sensed data, nongovernmetal organizations and monitors undertake patrols in forests |
| Frequency | Semimonthly volunteer monitoring | Weekly reports | Semimonthly visits to wells | Monthly patrols along transects | Satellite-based alerts; field patrols follow | Quarterly reporting |
| Aggregation/ analysis | Semimonthly information aggregated into quarterly reports and posters | Summary report generated by the smartphone app | Summary infographics generated by research teams | Poster based on monitoring information; Discussion of forest use (trends) in monthly community meeting of village households | Paper reports and digital records of monitoring effort and dynamics over time | Monitors share information from patrols (no aggregation except at end of experiment) |
| Information sharing | Local and provincial governments via quarterly reports, public and users via posters | Water managers and users | WhatsApp messages sent among water committee members, community members, and research team | Monthly community meetings of village households | Community, who then decides whether to share with government | Community members in group meeting after each patrol |

***Consequential Variation in the Treatment.*** All treatments are heterogeneous. Consider an experimental vaccine, which, at first glance, appears to be a simple, homogenous treatment. But its administration will likely vary in multiple dimensions. For example, the same vaccine can vary across individuals in ways such as potency, the quantity injected, and the timing of the injection. Most such variations are minor, sufficiently minor to be ignored when assessing efficacy of the vaccine. In contrast, the administration of an institutional intervention designed to affect the status and trajectory of a coupled human–natural system is both complex and heterogeneous: its variations may not be as easily ignored. How do we know when variation in the treatment is consequential?

We propose a mechanism-based, conceptual approach to understanding consequential variation in treatment. A treatment variation is consequential if different versions of the treatment create variations in mechanism effects. This approach to determining consequential variation is illustrated in Fig. 1.

In Fig. 1*A*, a version of a community monitoring-program, $C_{v1}$, has heterogeneous effects on resource status, *R*, conditional on-site conditions, *S*. This heterogeneity stems from moderating conditions at the sites that affect the channels, *M*, through which $C_{v1}$ affects *R*. For example, preexisting norms or institutions may moderate the degree to which management authorities perceive pressure to change their behaviors (i.e., these preexisting conditions moderate an accountability mechanism). The moderating conditions either affect how $C_{v1}$ affects *M* or how *M* affects *R*, or both (i.e., those causal pathways are represented in the figure by the arrows from treatment to mechanism and the arrows from mechanism to outcome). Fig. 1*A* thus illustrates what researchers mean by the phrase "heterogeneous treatment effects" or "conditional treatment effects"; that is, a community-monitoring program has different effects in different subgroups of the target population of communities (i.e., site conditions).[†]

Fig. 1*B* portrays a second version of the community-monitoring program, $C_{v2}$, which affects the mechanisms in the same ways that $C_{v1}$ affects them under the same site conditions. In this case, the variation in the two monitoring programs is inconsequential. In other words, a policymaker could expect a similar average effect on resource status from the two versions of the monitoring program.

Other versions of the treatment intervention, however, may not generate the same effects. In Fig. 1*C*, a third version of the community-monitoring program, $C_{v3}$, affects the mechanisms in the same way as $C_{v1}$ at site $S_1$, but not at site $S_2$. In this context, the moderating conditions at the site interact with the version of the monitoring program to create consequential variation. For example, whether a monitoring program has a mission statement that explicitly states the expected effects of the program on resource status may not matter at site $S_1$, where the population is

[†]The scale of the "site" in Fig. 1 is left undefined. One may be interested in how differences in conditions across communities in the same RCT may moderate the effect of *C* on *R* (within-study heterogeneity) or how differences in conditions across RCTs may moderate the effect of *C* on *R* (across-study heterogeneity).

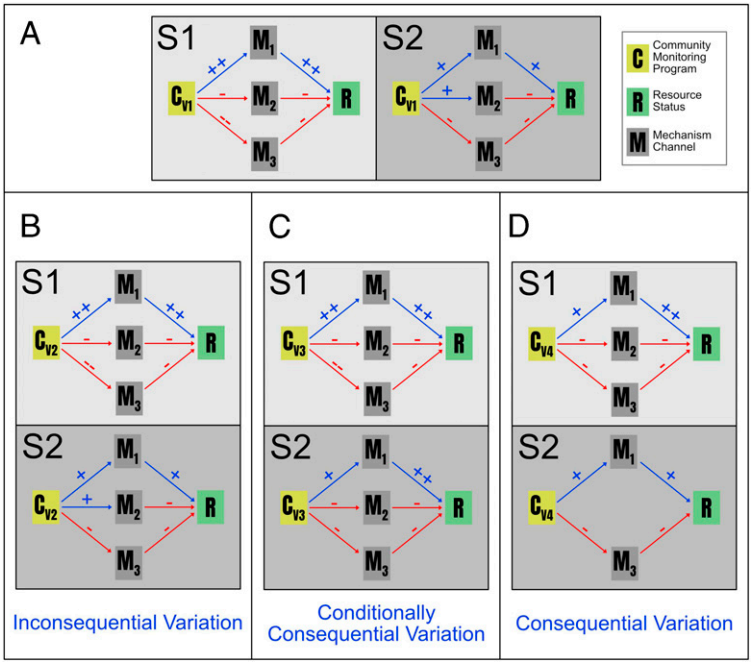


**Fig. 1. Consequential variation. A version of a community-monitoring program ($C_v$) can affect a resource status ($R$) through different mechanism channels ($M$) under different site conditions ($S$). Site conditions can moderate how $C$ affects $M$ and how $M$ affects $R$. Directed arrows indicate the path of causality and the (+) or (−) symbols indicate direction and magnitudes of the causal effects. (*A*) $C_{v1}$ (version 1) has heterogeneous treatment effects based on site conditions. The other panels are contrasted with this panel. (*B*) $C_{v2}$ affects $R$ in the same way that $C_{v1}$ affects $R$, and thus the treatment variation is inconsequential. (*C*) $C_{v3}$ affects $R$ in the same way that $C_{v1}$ affects $R$ at $S_1$, but not at $S_2$, and thus the variation is conditionally consequential (conditional on $S$). (*D*) At all sites, $C_{v4}$ affects $R$ in ways that differ from the ways that $C_{v1}$ affects $R$, and thus the variation is consequential. See text for more details.**

largely illiterate. Yet, at site $S_2$, the presence of a written mission statement changes the expectations of the players in the monitoring program (e.g., users, monitors, management authorities) in ways that affect the accountability channel. In other words, the variation is conditionally consequential, and a policymaker should not expect a similar effect from the two versions unless implementing both versions under similar conditions.

In Fig. 1*D*, a fourth version of the community-monitoring program, $C_{v4}$, affects the mechanisms differently at all sites. In fact, at site $S_2$, one of the mechanisms is not operative at all. This variation in the program attributes is thus consequential: it creates policy-relevant variation in the mechanism effects and thus in the overall effect of community monitoring on resource status.[‡]

In all four panels of Fig. 1, we see heterogeneity in the post-intervention status of the resource. The effect of a specific version of a community-monitoring program differs depending on where the version is implemented (i.e., the effects of a version are heterogeneous, also known as heterogeneous treatment effects). In Fig. 1 *C* and *D*, the effects of two versions of a community-monitoring program also differ at the same locations (i.e., the programs themselves are heterogeneous, also known as heterogeneous treatments).

This subtlety in terminology reflects a subtlety in interpretation when scientists describe heterogeneous outcomes within a single study or across studies. Does the heterogeneity arise from variation in moderating conditions across locations or from variation in the versions of the treatment implemented in the locations? Consequential variation in the treatment is a problem for scientists interested in causal inference, whether they rely on observational or experimental data. In other words, it is a threat to the internal validity of individual studies [when the variation is within-study (67)] and a threat to the internal validity of metaanalyses (66).

Consequential variation is also relevant to policy. Policymakers interested in designing interventions need to understand whether they are contemplating variations in the design of the intervention that may be consequential. From this perspective, the notion of consequential variation in the treatment design relates to the notion of external validity. Both notions are aspects of the broader notion of generalizability. When scientists refer to "generalizability," they typically mean external validity in the sense of whether the treatment, *C*, can be expected to affect the outcome, *R*, in the same way under different conditions. Scientists are not typically referring to generalizability across different versions of *C* itself. But because policy treatments can be highly heterogeneous, policy makers also want to know whether variants of *C* also affect *R*, and by how much and under what conditions. Unlike pharmaceutical treatments, for example, policy treatments aimed at affecting outcomes in coupled human–natural systems will typically vary across time and space. As we have seen, even a construct as seemingly simple as community monitoring can be implemented in myriad ways.[§]

[‡]One could imagine an indeterminate case, in which two versions affect *R* via different mechanisms or different magnitude effect sizes, but countervailing effects cancel each other out and the overall effect is the same. Given that such cases are unlikely in practice, we ignore them here.

[§]The source of exogenous variation in treatment exposure may also lead to consequential variation. For example, in an experimental design, the mechanisms may operate differently if the attributes of the agent offering the program send a signal to communities about the magnitude of expected benefits from the program and thus affects their investment in the program (e.g., differing signals from a local nonprofit, an international nonprofit, a local government agency, or a national government agency).

The metaketa teams' use of preregistered, randomized experimental designs made it less likely that their treatment interventions varied consequentially within each country. The metaketa harmonization process, with its emphasis on a common theory of change and similar field implementation, also made it less likely that the treatment interventions varied in consequential ways across countries. Despite these precautions, variation in the implemented version of the community monitoring did occur (e.g., in China, monitors did not send monitoring reports to management authorities, thus potentially weakening the accountability channel). The authors and their readers are thus forced to make untested (and potentially untestable) assumptions about the consequentiality of treatment variations within and across countries.

How does one judge whether a variation is consequential? Our definition and elaboration of consequential variation makes clear that such judgement is impossible in the absence of an elaborated, mechanism-based model of how the treatment can affect an outcome. Such models are clearly necessary to guide judgements about consequential variation in science, when aiming to infer causality within a setting or to generalize causal relationships across settings. But they are also critical to policy, when aiming to determine whether an anticipated variation in the program design will be consequential. Neither empirical science nor policymaking can afford to be atheoretical.[¶]

Assessing consequential variation through the lens of elaborate, mechanism-based theories can help in the design of single studies and in harmonization across studies. The study teams in this Special Feature used a harmonized, mechanism-based theory of change, which they used to judge whether proposed treatment variations would be consequential. This theory also guided the teams' selection of relevant intermediate variables and their relationships (mechanisms), allowing the teams to assess whether the treatment was working in similar ways in the six countries. Based on analyses of these intermediate variables, the accountability channel seemed to be most influential in this set of studies, suggesting that the variations in treatment design were not consequential unless they affected the accountability channel (the smallest treatment effect on resource use was in the China context, where monitoring reports did not go directly to authorities). This mechanism analysis at the level of individual studies was complemented by analysis of heterogeneity in the metaanalysis (14). The authors found no statistical evidence of heterogeneity in the effect sizes from the six country studies, a finding that further strengthens the assumption of no consequential variation in treatment across the studies (however, given there are only six studies, the metaanalysis is only powered to detect substantial heterogeneity). In other words, readers may plausibly assume that the metaketa teams have implemented six inconsequential variations of the treatment, each assigned to a different country in 35 to 80 communities per country.

If a metaanalysis detects evidence of heterogeneity in the estimated effect sizes across the studies, one cannot necessarily infer that the treatments vary in consequential ways. This is because although metaanalyses can provide statistical evidence of heterogeneity in treatment effects across studies, they cannot easily identify the sources of heterogeneity. Heterogeneity across studies can arise from random error, heterogeneous treatment effects (Fig. 1*A*), consequential variation in treatments (Fig. 1 *B* and *C*), different outcome measures, or different designs and methods. Without an elaborate, mechanism-based theory, researchers cannot credibly distinguish among these explanations. For example, the metaketa used similar designs and methods and thus could rule out those explanations, but each context had a different resource status measure. The explanation that treatment effects vary by resource status can only be ruled out if the team is correct that the theory of change is independent of the resource: the mechanisms would work in the same way for groundwater as for a forest. If the metaanalysis had produced evidence of heterogeneity, the authors would thus have to determine whether that heterogeneity arose because of variations in the study contexts or because of consequential variations in the treatment, or both. The authors could have chosen to simply make an assertion (e.g., ruling out consequential variations in treatment based on the harmonization process) or to run an empirical analysis (e.g., metaregression) that identifies potentially consequential features of the contexts and treatments and estimates how each feature contributes to variation in the estimated effect sizes (an approach that typically requires more studies than six). Yet, to be credible, assertions and choices of features to include in an empirical analysis can only be justified when based on an elaborate, mechanism-based theory. In fact, such theory is required for any attempt to synthesize empirical data, regardless of whether the synthesis is done through a meta-analysis or other comparative approach.

Thus, sustainability science is going to be most successful when it is practiced as an iterative process between theory development and empirical analyses, whereby theory-driven hypotheses about what variations are consequential are first posed and then tested empirically. The constraints to knowledge accumulation posed by consequential variation exist in any empirical study of sustainability interventions, whether the study is experimental or nonexperimental or whether it is preregistered and harmonized. Ultimately, in science, assessing whether a treatment variation is consequential requires shared assumptions about the underlying model that characterizes the system being studied. This set of shared assumptions allows scholars to determine whether a theory is sufficiently deep: that is, whether the mechanisms are sufficiently elaborated at their lowest level to assess if a treatment variation is consequential.

***The Metaketa Design and Implications for Accumulating Knowledge.*** All causal effects are defined as a contrast between two states of the world. What are the two states of the world being contrasted in the RCTs in this Special Feature? In the simplest terms, the RCTs measure differences in outcomes between the status quo and a state in which a community monitoring intervention is added to the status quo. However, the status quo in these RCTs is not "the absence of any community monitoring." The communities in which the monitoring programs were assigned had some monitoring prior to the RCT intervention (see *SI Appendix*, Section S6 in ref. 14). In fact, it's hard to imagine any community that uses a common pool resource but has zero community monitoring. For example, even without formal, specialized monitoring roles, community members observe each other while carrying out their daily routines. Community monitoring is thus best viewed as a continuous construct that varies by the intensity or extent of the monitoring.

[¶]The importance of mechanistic theorizing (also known as the elaboration of generative mechanisms) to support the interpretation of causal analyses has been long asserted by social scientists (e.g., refs. 68 and 69), statisticians (e.g., ref. 70), and philosophers of science (e.g., refs. 71 and 72); see summary in chapter 10 of ref. 73.

From this perspective, the studies in the Special Feature contrast "enhanced community monitoring" (treatment) to "status quo community monitoring" (control). The estimated effects in the RCTs may thus best be viewed as lower-bound estimates of the expected effects of adding a maximal community-monitoring program to a condition of no monitoring at the same study sites (unless the function that maps doses of monitoring intensity to outcomes is nonmonotonic).

We can further unpackage the construct of enhanced community monitoring. A community-monitoring program can be enhanced via internal forces (i.e., endogenously) or via external forces (i.e., exogenously). In the common pool resource literature, most community-monitoring programs seem to arise endogenously: that is, the communities that are most likely to adopt enhanced monitoring systems are the ones that perceive the most gain from developing them and, perhaps, have external agents to assist them in setting up the programs. That endogenous selection into monitoring makes studying the effects of monitoring a challenge in nonexperimental designs. The challenge arises because the communities that adopt monitoring are different from the communities that do not adopt monitoring in ways that may also affect resource conditions and community welfare, making it difficult to estimate what would have happened in the monitored communities had they not had monitoring (i.e., counterfactual outcomes).

The metaketa in this Special Feature solves this endogenous selection challenge by using an experimental design. The teams randomized their monitoring interventions within a group of communities recruited from a population of communities that had not yet implemented a formalized monitoring program but were interested in implementing such a program.

However, this solution to the endogenous selection challenge affects the nature of the causal effect that the teams can estimate. The teams estimate the expected effect of an "external actor-initiated" enhancement of community monitoring, rather than of a community-initiated enhancement of monitoring. Furthermore, they estimate this causal effect in a subgroup of communities. This subgroup comprises communities that are willing to accept external support for establishing an externally designed, enhanced community-monitoring system. This subgroup excludes communities that already have an enhanced system of monitoring that they designed and adopted on their own (i.e., the focus on much of the common pool resource management literature) and communities that express no interest in adopting such a monitoring system (i.e., part of the population that policymakers care most about).

Thus, without more assumptions, the RCTs in the Special Feature do not shed light on the expected impact of endogenously generated community monitoring or the expected impact of community monitoring in communities that are not interested in adopting monitoring without more incentives or coercion. One could extrapolate to those groups if, for example, one is willing to assume that all communities have good information about the returns to monitoring. In that case, one could assume that the average "already have it" community would have an average treatment effect (ATE) higher than the ATE for the average "don't have it but want it" community. The "don't have it but want it" community would, in turn, have an ATE higher than the average "don't have it and don't want it" community (i.e., if that community were forced to have a monitoring program via, for example, a government rule or regulation). But without assumptions like these, such extrapolations are not possible.

Nevertheless, from a policy perspective, the effect in the "don't have it but want it" communities may be the most relevant effect. Viewed from this perspective, the study contexts in the Special Feature may be more relevant to the decentralization, comanagement, and state-recognized community-based management contexts (74, 75), and perhaps less relevant to contexts where community monitoring emerges endogenously, as was the case with the original common pool research studies included in (1).

Further unpacking the construct of community monitoring in these trials, one ought to consider how monitors were incentivized in these trials. Monitors were paid conditional on submitting their reports because the study teams wanted to reduce the likelihood of noncompliance by the monitors. In other words, the RCTs were designed to ensure that monitoring took place. One could thus think of these RCTs as "efficacy trials," in which atypical efforts are taken to enhance compliance, rather than "effectiveness trials," in which compliance may be more variable in space and time under so-called "natural" conditions. Understanding the first-order problem of whether a monitoring program can affect resource outcomes is important, but that understanding does not provide guidance on solving the second-order problem of how to motivate monitors without outside aid or how outside aid from different sources may lead to different rates of compliance (e.g., coming from a government rather than nonprofit organization). Solving that second-order problem is a separate research agenda.

Finally, we want to highlight the limitations of using short-term interventions to draw inferences in sustainability science. The postintervention period in the six RCTs is roughly 1 y. Sustainable common pool resource management, however, requires persistent improvement in management. The estimated average effects reported in the Special Feature are likely to differ from the average effects of a persistent change in community monitoring. But predicting how they may differ depends on assumptions about long-term mechanisms. The estimates in the Special Feature could be lower-bound estimates of a persistent change in community monitoring if there were learning-by-doing (maturation) within the monitoring programs or if norms change slowly over time. However, the same estimates could be upper-bound estimates if shirking of monitoring effort (free riding) increases over time. Moreover, the RCTs shed no light on the question of how to institutionalize the monitoring programs into the broader governance processes in civil society.

## The Future of Common Pool Resource Management and Empirical Sustainability Science

Despite the limitations described in the previous section, this Special Feature makes three contributions to advance sustainability science. The first is an assessment of community monitoring from the perspective of four sustainability science literatures. The second contribution is an exemplar of a research design, the metaketa, that can strengthen empirical sustainability science. The third contribution lies in an illustration of the challenges for sustainability scientists hoping to synthesize prior research to better understand causal relationships in socio-ecological systems.

Scholarship on community monitoring in common pool resource systems reaches divergent conclusions about whether community monitoring is effective only when it emerges endogenously in a community, or whether it can also be effective when introduced by external agents, such as a central government. Related to this argument is whether community monitoring is effective as a standalone intervention or only when it is nested in a set of institutional features that include sanctions on rule breakers and arrangements for adjudicating disputes. Intervening in this debate, the studies in this special issue show that institutional

design features, such as community monitoring, can be manipulated individually in common pool resource settings, and that communities are willing to adopt enhanced monitoring interventions when accompanied by technical and financial support (but see ref. 11).

This empirical evidence is important because prior theory and empirical analyses offer little guidance to scholars or policymakers on whether the dozens of identified design principles for effective resource governance have to be enhanced simultaneously to change outcomes, or if they can be manipulated independently. Additionally, the fact that community monitoring can be introduced as an individual institutional intervention also means that the intervention can be scaled across communities whose institutions may differ from each other, making it relevant across a larger set of literatures in sustainability science: for example, on decentralization, community-driven development, or payments for ecosystem services (76). Indeed, one benefit of harmonized, experimental designs is that they require the treatments to be scalable. However, the metaketa studies do not provide evidence on whether one should expect additive impacts from multiple changes in institutional design features. Future studies could explore whether multiple changes to institutional design features have additive or multiplicative effects, are substitutes for one another, or worse, interfere with each other. The metaketa highlights how large—and thus expensive—such studies will be to have sufficient power to detect policy-relevant differences across multiple treatment arms (although adaptive designs may lower that cost).

Knowing that an institutional intervention like community monitoring can be deployed individually is important but insufficient. For both scientific advances and policy actions, knowing if the intervention has an effect on outcomes is also important. The second major finding of this Special Feature is that, on average, externally driven enhancements of community monitoring in social-ecological systems cause reductions in resource extraction and increases in user satisfaction [however, see cautions on interpretation expressed by Barrett (16)]. In addition, the meta-analysis of the six studies detects no statistical evidence of heterogeneous effects (14). This lack of statistical evidence, in combination with the theory and evidence about the operative mechanisms (see below), suggests that the variations in monitoring programs across the six countries were not consequential and that the programs had comparable effects despite being implemented in different contexts.

The summary estimated effect size on reductions in resource extraction is 0.10 SDs (14), which is comparable to the summary estimated effect sizes in metaanalyses of interventions in other sustainability settings: payments for environmental services (77), information-based strategies in energy conservation (78), and climate change mitigation (79). It is also comparable to the magnitudes of estimated effect sizes in evaluations of protected areas that address the nonrandom location of the areas (80). An effect size of 0.10 SD across a range of interventions has an important implication for sustainability scientists: they need to think more about statistical power. Only one of the studies in the Special Feature was designed to detect an effect of 0.10 SD with power near the conventional level of 80%. Had we only had one or a few of the six studies, we may have failed to detect the effect of monitoring. Perhaps worse, if only the Amazonian study that had the largest estimated effect size was published (13), researchers and policymakers may have interpreted community monitoring as much more impactful than it likely is [the problem of exaggeration biases in literatures with underpowered designs and publication biases against statistically or scientifically "insignificant" results is slowly being understood across disciplines (81)]. The problem of low statistical power is relevant beyond experimental designs. Multicountry, nonexperimental studies of community-based natural resource management also tend to base their findings on a relatively small number of cases, with only a couple of hundred observations or fewer being common (e.g., refs. 82–85).

The third major insight related to community monitoring is that the effects of community monitoring on resource status and extraction appear to be mediated primarily through accountability channels. This insight has two implications. First, monitoring programs implemented in other common pool resource contexts are most likely to have similar effects to the programs in the Special Feature when those contexts include management authorities whose behaviors have important effects on resource status and user satisfaction. Second, variations in program design are most likely to be consequential when they affect accountability channels. For example, the way in which monitoring information is collected and presented to other users (e.g., through remote sensing or in-person patrols, digitally or paper-and-pen) may not matter as much as ensuring the information is delivered to the relevant management authorities. Guidance on consequential variation in the treatment is important because, for institutional treatments relevant to sustainability, empirically assessing the consequentiality of potential variations in a treatment design would be costly.

Nevertheless, despite the important insights provided by the metaketa, it is worth noting that institutions are complex, measures of institutional features remain in their infancy, and knowledge about how features of institutions interact with each other and with their resource, socioeconomic, and political contexts is inadequate at best. While the metaketa provides systematic evidence on community monitoring, practitioners interested in changing institutions need also to rely on a mix of: 1) prior evidence; 2) a shared, elaborated, mechanism-based theory of the relationships among the treatment, the units of analysis, intermediate outcomes, and contextual covariates; and 3) a commitment to pursuing new evidence where it is missing. Ultimately, generalizability requires that scientists uncover not only the average effects of changes in institutional features of systems, but also whether institutional effects work through the mechanisms in the ways scholars and practitioners hypothesize they do.

In addition to insights related to community monitoring, the second major contribution of the Special Feature lies in the example it offers for a novel approach to empirical research in sustainability science to improve causal inference. Although randomized controlled trials and replications are rare in the sustainability science literature on coupled human–natural systems, they do exist (e.g., refs. 86–92). But preregistered, harmonized replications across heterogeneous contexts are absent. Such replications enhance the reliability of research findings by emphasizing common research designs (with an emphasis on comparable causes, theories of change, and outcomes), constraints on researcher degrees of freedom (through preregistration of implementation and analysis plans and third-party analyses), synthesis across studies, and coordinated publications. Indeed, these principles are relevant to the conduct of nonexperimental studies as well. They promise to yield substantial benefits for advancing empirical sustainability science.

The third contribution of the Special Feature concerns how it highlights the challenges faced by researchers who seek to synthesize heterogeneous research findings and by policymakers who seek to apply the research findings in heterogeneous circumstances. Precisely because institutions and social-ecological systems are complex, heterogeneity will likely be the rule rather

than the exception. For sustainability science to fulfill its promise, sustainability scientists and policymakers need to understand the driving forces and implications of this heterogeneity. Although this understanding can be informed by *ex ante* efforts to limit the variation of causes and contexts under study and *ex post* efforts to statistically assess the importance of different driving forces, these efforts are expensive: they require estimates of causal effects from multiple iterations of different versions of a causal factor and different contexts in which the factor may operate. In this introductory article, we build on prior work to elaborate a mechanism-based framework for assessing consequential variation in causes and contexts. When scientists and policymakers regularly propose elaborate, mechanism-based theories through which their proposed causes have effects, the prospects for more credible synthesis and more successful field applications in sustainability science will grow.


## Acknowledgments

For constructive comments on the manuscript, we thank Paul Feldman, Jordi Honey-Roses, Inés Ibáñez, Christoph Nolte, Pallavi Shukla, Tara Slough, Cyrus Samii, and Billie Turner. P.J.F. thanks K. Rowles for assistance in creating Fig. 1. A.A. thanks Maria Lemos for discussions on the topics in this introductory article.



1 E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, 1990).
2 T. Dietz, E. Ostrom, P. C. Stern, The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).
3 F. Saunders, The promise of common pool resource theory and the reality of commons projects. *Int. J. Commons* **8**, 636–656 (2014).
4 J. Baggio *et al.*, Explaining success and failure in the commons: The configural nature of Ostrom's institutional design principles. *Int. J. Commons* **10**, 417–439 (2016).
5 A. Agrawal, Common property institutions and sustainable governance of resources. *World Dev.* **29**, 1649–1672 (2001).
6 P. Olsson, C. Folke, F. Berkes, Adaptive comanagement for building resilience in social-ecological systems. *Environ. Manage.* **34**, 75–90 (2004).
7 E. Ostrom, *Understanding Institutional Diversity* (Princeton University Press, 2009).
8 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).
9 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).
10 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).
11 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).
12 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).
13 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).
14 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).
15 E. Asiedu, D. Karlan, M. Lambon-Quayefio, C. Udry, A call for structured ethics appendices in social science papers. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024570118 (2021).
16 C. B. Barrett, On design-based empirical research and its interpretation and ethics in sustainability science. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023343118 (2021).
17 W. C. Clark, A. G. Harley, Sustainability science: Toward a synthesis. *Annu. Rev. Environ. Resour.* **45**, 331–386 (2020).
18 P. J. Ferraro, J. N. Sanchirico, M. D. Smith, Causal inference in coupled human and natural systems. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5311–5318 (2019).
19 C. J. Bryan, D. S. Yeager, J. M. O'Brien, Replicator degrees of freedom allow publication of misleading failures to replicate. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 25535–25545 (2019).
20 J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
21 D. Fanelli, R. Costas, J. P. Ioannidis, Meta-assessment of bias in science. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3714–3719 (2017).
22 J. D. West, C. T. Bergstrom, Misinformation in and about science. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1912444117 (2021).
23 E. T. Borer *et al.*, Finding generality in ecology: A model for globally distributed experiments. *Methods Ecol. Evol.* **5**, 65–73 (2014).
24 L. H. Fraser *et al.*, Coordinated distributed experiments: An emerging tool for testing global hypotheses in ecology and environmental science. *Front. Ecol. Environ.* **11**, 147–155 (2013).
25 A. Banerjee *et al.*, Development economics. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science* 348, 1260799 (2015).
26 A. Przeworski, H. Teune, *The Logic of Comparative Social Inquiry* (John Wiley and Sons, Inc., New York, 1970).
27 J. Seawright, J. Gerring, Case selection techniques in case study research: A menu of qualitative and quantitative options. *Polit. Res. Q.* **61**, 294–308 (2008).
28 T. Dunning, G. Grossman, M. Humphreys, S. D. Hyde, C. M. McIntosh, "The metaketa initiative" in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, T. Dunning *et al.*, Eds. (Cambridge Studies in Comparative Politics, Cambridge University Press, Cambridge, 2019), pp. 16–49.
29 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).
30 A. Deaton, Instruments, randomization, and learning about development. *J. Econ. Lit.* **48**, 424–455 (2010).
31 R. McDermott, P. K. Hatemi, Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30014–30021 (2020).
32 R. M. Kaplan, V. L. Irvin, Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One* **10**, e0132382 (2015).
33 B. A. Nosek, C. R. Ebersole, A. C. DeHaven, D. T. Mellor, The preregistration revolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2600–2606 (2018).
34 C. F. Manski, Policy analysis with incredible certitude. *Econ. J. (Lond.)* **121**, F261–F289 (2011).
35 C. Pahl-Wostl, A conceptual framework for analysing adaptive capacity and multi-level learning processes in resource governance regimes. *Glob. Environ. Change* **19**, 354–365 (2009).
36 C. Stem, R. Margoluis, N. Salafsky, M. Brown, Monitoring and evaluation in conservation: A review of trends and approaches. *Conserv. Biol.* **19**, 295–309 (2005).
37 G. M. Lovett *et al.*, Who needs environmental monitoring? *Front. Ecol. Environ.* **5**, 253–260 (2007).
38 J. L. Dickinson *et al.*, The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.* **10**, 291–297 (2012).
39 D. Couvet, F. Jiguet, R. Julliard, H. Levrel, A. Teyssedre, Enhancing citizen contributions to biodiversity science and public policy. *Interdiscip. Sci. Rev.* **33**, 95–103 (2008).
40 F. Danielsen *et al.*, Local participation in natural resource monitoring: A characterization of approaches. *Conserv. Biol.* **23**, 31–42 (2009).
41 A. I. Tulloch, H. P. Possingham, L. N. Joseph, J. Szabo, T. G. Martin, Realising the full potential of citizen science monitoring programs. *Biol. Conserv.* **165**, 128–138 (2013).

**42** F. Danielsen *et al.*, The concept, practice, application, and results of locally based monitoring of the environment. *Bioscience* **71**, 484–502 (2021).
**43** National Academies of Sciences, Engineering, and Medicine, *Learning Through Citizen Science: Enhancing Opportunities by Design* (NAS, Washington, DC, 2018).
**44** C. C. Conrad, K. G. Hilchey, A review of citizen science and community-based environmental monitoring: Issues and opportunities. *Environ. Monit. Assess.* **176**, 273–291 (2011).
**45** S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, Eds., *Citizen Science: Innovation in Open Science, Society and Policy* (University College London Press) 2018).
**46** F. Berkes, J. Colding, C. Folke, Rediscovery of traditional ecological knowledge as adaptive management. *Ecol. Appl.* **10**, 1251–1262 (2000).
**47** K. A. Waylen *et al.*, Policy-driven monitoring and evaluation: Does it support adaptive management of socio-ecological systems? *Sci. Total Environ.* **662**, 373–384 (2019).
**48** D. A. Keith, T. G. Martin, E. McDonald-Madden, C. Walters, Uncertainty and adaptive management for biodiversity conservation. *Biol. Conserv.* **4**, 1175–1178 (2011).
**49** E. S. G. Schreiber, A. R. Bearlin, S. J. Nicol, C. R. Todd, Adaptive management: A synthesis of current understanding and effective application. *Ecol. Manage. Restor.* **5**, 177–182 (2004).
**50** M. Fujitani, A. McFall, C. Randler, R. Arlinghaus, Participatory adaptive management leads to environmental learning outcomes extending beyond the sphere of science. *Sci. Adv.* **3**, e1602516 (2017).
**51** D. Rustagi, S. Engel, M. Kosfeld, Conditional cooperation and costly monitoring explain success in forest commons management. *Science* **330**, 961–965 (2010).
**52** D. Rahman, But who will monitor the monitor? *Am. Econ. Rev.* **102**, 2767–2797 (2012).
**53** J. I. Ricks, Building participatory organizations for common pool resource management: Water user group promotion in Indonesia. *World Dev.* **77**, 34–47 (2016).
**54** M. A. Rudd, An institutional framework for designing and monitoring ecosystem-based fisheries management policy experiments. *Ecol. Econ.* **48**, 109–124 (2004).
**55** R. J. Barro, The control of politicians: An economic model. *Public Choice* **14**, 19–42 (1973).
**56** C. Lessmann, G. Markwardt, One size fits all? Decentralization, corruption, and the monitoring of bureaucrats. *World Dev.* **38**, 631–646 (2010).
**57** D. Mookherjee, Political decentralization. *Economics* **7**, 231–249 (2015).
**58** T. Campbell, *The Quiet Revolution: The Rise of Political Participation and Leading Cities with Decentralization in Latin America and the Caribbean* (University of Pittsburgh Press, Pittsburgh, 2001).
**59** J. A. Rodden, J. M. Rodden, *Hamilton's Paradox: The Promise and Peril of Fiscal Federalism* (Cambridge University Press, Cambridge, 2006).
**60** A. Mejía Acosta, The impact and effectiveness of accountability and transparency initiatives: The governance of natural resources. *Dev. Policy Rev.* **31**, s89–s105 (2013).
**61** A. Joshi, Legal empowerment and social accountability: Complementary strategies toward rights-based development in health? *World Dev.* **99**, 160–172 (2017).
**62** B. A. Olken, Monitoring corruption: Evidence from a field experiment in Indonesia. *J. Polit. Econ.* **115**, 200–249 (2007).
**63** J. A. Fox, Social accountability: What does the evidence really say? *World Dev.* **72**, 346–361 (2015).
**64** E. Molina, L. Carella, A. Pacheco, G. Cruces, L. Gasparini, Community monitoring interventions to curb corruption and increase access and quality in service delivery: A systematic review. *J. Dev. Effect.* **9**, 462–499 (2017).
**65** H. Waddington *et al.*, Citizen engagement in public services in low-and middle-income countries: A mixed-methods systematic review of participation, inclusion, transparency and accountability (PITA) initiatives. *Campbell Syst. Rev.* **15**, e1025 (2019).
**66** S. G. Thompson, Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* **309**, 1351–1355 (1994).
**67** T. J. VanderWeele, M. A. Hernán, Causal inference under multiple versions of treatment. *J. Causal Inference* **1**, 1–20 (2013).
**68** J. H. Goldthorpe, Causation, statistics, and sociology. *Eur. Sociol. Rev.* **17**, 1–20 (2001).
**69** P. Hedström, R. Swedberg, Eds., *Social Mechanisms: An Analytical Approach to Social Theory* (Cambridge University Press, Cambridge, UK, 1998).
**70** P. Rosenbaum, *Observational Studies* (Springer, New York, NY), ed. 2, 2002).
**71** M. Bunge, How does it work? The search for explanatory mechanisms. *Philos. Soc. Sci.* **34**, 182–210 (2004).
**72** P. Machamer, L. Darden, C. F. Craver, Thinking about mechanisms. *Philos. Sci.* **67**, 1–25 (2000).
**73** S. L. Morgan, C. Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (Cambridge University Press, New York, NY, 2015) ed. 2.
**74** J. E. Cinner, *et al.*, Comanagement of coral reef social-ecological systems. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5219–5222 (2012).
**75** J. C. Ribot, "Democratic decentralization of natural resources" in *Beyond Structural Adjustment: The Institutional Context of African Development* (Palgrave Macmillan, New York, 2003) pp. 159–182.
**76** C. Samii, M. Lisiecki, P. Kulkarni, L. Paler, L. Chavis, Effects of decentralized forest management (DFM) on deforestation and poverty in low and middle income countries: A systematic review. *Campbell Syst. Rev.* **10**, 1–88 (2014).
**77** B. Snilstveit *et al.*, Incentives for climate mitigation in the land use sector: A mixed-methods systematic review of the effects of payment for environmental services (PES) on environmental and socio-economic outcomes in low- and middle-income countries. *Campbell Systemic Reviews* **15**, e1045 (2019).
**78** M. Delmas, M. Fischlein, O. I. Asensio, Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. *Energy Policy* **61**, 729–739 (2013).
**79** C. F. Nisa, J. J. Bélanger, B. M. Schumpe, D. G. Faller, Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nat. Commun.* **10**, 4545 (2019).
**80** P. J. Ferraro *et al.*, Estimating the impacts of conservation on ecosystem services and poverty by integrating modeling and evaluation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7420–7425 (2015).
**81** A. Gelman, J. Carlin, Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).
**82** A. Chhatre, A. Agrawal, Forest commons and local enforcement. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13286–13291 (2008).
**83** A. Chhatre, A. Agrawal, Trade-offs and synergies between carbon storage and livelihood benefits from forest commons. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17667–17670 (2009).
**84** L. Persha, A. Agrawal, A. Chhatre, Social and ecological synergy: Local rulemaking, forest livelihoods, and biodiversity conservation. *Science* **331**, 1606–1608 (2011).
**85** R. Rasolofoson *et al.*, Impacts of community forest management on human economic well-being in Madagascar. *Conserv. Lett.* **10**, 346–353 (2016).
**86** M. A. Andor, A. Gerster, J. Peters, C. M. Schmidt, Social norms and energy conservation beyond the US. *J. Environ. Econ. Manage.* **103**, 102351 (2020).
**87** A. Brandon *et al.*, (2017). Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments. *NBER Working Paper Series*. https://www.nber.org/papers/w23277. Accessed 25 February 2021.
**88** D. A. Brent, J. H. Cook, S. Olsen, Social comparisons, household water use, and participation in utility conservation programs: Evidence from three randomized trials. *J. Assoc. Environ. Resour. Econ.* **2**, 597–627 (2015).
**89** P. J. Ferraro, M. Price, Using non-pecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Rev. Econ. Stat.* **95**, 64–73 (2013).
**90** S. Jayachandran *et al.*, Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science* **357**, 267–273 (2017).
**91** E. Wiik *et al.*, Experimental evaluation of the impact of a payment for environmental services program on deforestation. *Conserv. Sci. Pract.* **1**, e8 (2019).
**92** B. Wilebore, M. Voors, E. H. Bulte, D. Coomes, A. Kontoleon, Unconditional transfers and tropical forest conservation: Evidence from a randomized control trial in Sierra Leone. *Am. J. Agric. Econ.* **101**, 894–918 (2019).