PART III:

Selecting, Measuring, Monitoring, and Evaluating Behavior Change Indicators

Part III of III from:

Behavior Change Interventions in Practice:

A synthesis of criteria, approaches, case studies & indicators

STAP SCIENTIFIC AND TECHNICAL ADVISORY PANEL
An independent group of scientists that advises the Global Environment Facility

gef UN environment programme

rare CENTER FOR BEHAVIOR & THE ENVIRONMENT

Authors:

Katie Williamson
Senior Associate, Center for Behavior & the Environment, Rare

Philipe M. Bujold
Senior Associate, Center for Behavior & the Environment, Rare

Erik Thulin
Behavioral Science Lead, Center for Behavior & the Environment, Rare

Recommended citation:

Williamson, K., Bujold, P. M., & Thulin, E. (2020). Behavior Change Interventions in Practice: A synthesis of criteria, approaches, case studies & indicators. Rare Center for Behavior & the Environment and the Scientific and Technical Advisory Panel to the Global Environment Facility.

Acknowledgments:

We would like to acknowledge the valuable review and comments on this report from Edward Carr (STAP), Graciela Metternicht (STAP), Mark Stafford Smith (STAP), Guadalupe Duron (STAP secretariat), Christopher Whaley (STAP secretariat), and Kevin Green (Rare BE.Center); the support of Andrea Wilk (Rare BE.Center) and Camille Freeman (Rare BE.Center) in case study development; research by Milan Urbanik (London School of Economics) and Ganga Shreedhar (London School of Economics) identifying behavior change frameworks and case studies; and Corinn Weiler (Rare) and Kyla Timberlake (Rare) for graphics development and document design.
Cover photo by George Stoyle.

This report was commissioned and funded by the Scientific and Technical Advisory Panel to the Global Environment Facility.

This work is licensed under CC BY 4.0.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0

Table of Contents

Part III: Selecting, Measuring, Monitoring, and Evaluating Behavior Change Indicators 4

Indicator Selection 5

Indicator Measurement 9

Framework for Indicator Monitoring and Adaptive Management 13

Frameworks for Evaluating Changes in Indicators 18

Cross-Context Generalizability 22

Conclusion: Recommendations On Indicators For Behavior Change Programming 23

References 24

Glossary 26

Introduction

Not only is the monitoring and evaluation of program indicators critical for accountability, it is also required if implementers are to improve program development, provide rapid data for decision making, and to allow for the generalization of findings to new and varying contexts. Despite the large majority of environmental programs implicitly requiring behavior change, many programs fail to include behavioral outputs as key indicators. It is even rarer for programs to systematically evaluate psychological and social indicators. This leaves programs with both poor metrics of success and an even poorer understanding of the underlying reasons for why a program achieved the level of success it did, and whether it will generalize to novel contexts.

This section is broken into four components. In the **first component**, we provide guidance on how to select behavioral, psychological, and social indicators. Since different programs tackle different challenges in different contexts, we cannot make general statements about the specific indicators that should be included. Instead, we provide guidance on how indicator selection can be based on a program's psycho-social theory of change. This, for example, includes the explicit representation of how program activities affect the psychological and social characteristics of a community, which then affects behavior and achieves program outcomes.

The **second component** focuses on how psycho-social and behavioral indicators can be measured. This includes a discussion of direct outcome observations, an option available for some 'publicly accessible' behaviors; proxy measures, which can indirectly assess an indicator; and self-reports, which are often necessary to measure psychological and social indicators.

The **third and fourth components** discuss how the monitoring and evaluation of psycho-social and behavioral indicators can improve program development. The monitoring of key psycho-social indicators allows not only for adaptive management, where program design is improved over time (based on insights into the program's mechanism) but also for what we term dynamic programming, where live programmatic decisions about phase transitions, expansion, or termination can be made based on real-time monitoring of psychological and social states within a target actor. Moving to evaluation, the quantification of program outcomes, including changes in target behaviors, allows program designers to assess program efficacy and effectiveness in order to make informed decisions over the scaling of a program. Finally, we conclude with a discussion of how the integration of monitoring data of psycho-social indicators with outcome evaluation allows for informed decisions about the generalizability of programs to novel contexts.

Indicator Selection

Selecting indicators is a critical component for developing the monitoring and evaluation (M&E) plan of any program—behavioral change efforts are no exception. Traditional M&E frameworks in the environmental field tend to track program delivery components for the monitoring and evaluation of environmental impact. Tracking program delivery and environmental impact is critical for successful M&E. In this section, we aim to expand from these two types of indicators and provide guidance on additional monitoring and evaluation components that are relevant to programs targeting behavior change.

A Psycho-Social Theory of Change

Emerging in the mid-90s, *Theories of Change* (ToC) have become the dominant paradigm for program planning and monitoring (Coryn et al., 2010). At its core, a program's ToC is a directional flow diagram that illustrates the relationship between program elements, intermediate outputs, and program outcomes. The elegance of the ToC paradigm is its explicitness: it forces a program designer to state each and every intermediary causal hypothesis of the program. In this way, having a ToC directs the development of the indicators to be measured: first, monitor the delivery of program elements; then, ensure that intermediary outputs have been achieved; finally, evaluate program outcomes. When well-executed, this type of monitoring allows designers to answer questions about why a program might not be performing optimally and provides them with the feedback needed for adaptive management.

While these elements of traditional ToCs and their accompanying indicators are critical, we propose that they are insufficient. This is because they implicitly assume critical psychological, social, and behavioral changes, without explicitly representing those changes within the ToC. However, the success of behavior change program development is entirely dependent on a proper understanding of the psychological and social environment in which an actor is behaving, and, more critically, on an understanding of how programming may change these environments or states. It is through changes to these states that we expect changes in behavior, which may, in turn, cause further changes to the social context. This means that quality behavior change programs based on that understanding of the psycho-social context include hypotheses about changes to the psychological and social characteristics of actors as well as their communities, including how interventions are expected to change these states, how these states affect behavior, and how that change in behavior cycles back to further change the social context. It is thus critical that the intermediate psychological and social indicators also be monitored. When measured properly, these indicators provide the same monitoring value as those in a traditional ToC, but also provide guidance for both diagnosing the cause of program success or failure, as well as facilitating adaptive management.

We employ the term *Psycho-Social Theory of Change* (PS-ToC) to describe the amalgamation of the above elements and more traditional ToCs. In a PS-ToC, between each element of an intervention and its intermediary behavioral outcomes lies the psychological or social changes assumed to have resulted from that intervention—i.e., the change that ultimately led to the behavioral outcome. It also allows for the representation of the dynamic relationship between behavior and psychological and social states, in which psychological and social states cause particular behaviors, which can, in turn, cause changes in those same psychological and social states. This explicit representation allows for a far greater understanding of why a program might be working or failing in terms of changes in the target psycho-social characteristics. It also allows for rapid adaptive management based on changes to those psycho-social factors long before an intermediate outcome might otherwise be observable. How this can be applied is discussed in detail in the *Frameworks for Indicator Monitoring and Adaptive Programming* section.

The application of a Psycho-Social Theory of Change requires the monitoring of three broad classes of indicators: behavioral, psychological, and social.

Behavioral Indicators

Monitoring changes in behavior is critical for any program claiming to target behavior change. In fact, for those behavior change programs where the link between the target behavior and environmental outcomes is already well documented, behavior may be evaluated as a final program outcome rather than an intermediate output measured in tandem with the environmental result of the intervention. Treating behavior as the program outcome is a particularly attractive option when a specific behavior is known to cause a longer-term environmental outcome, or when the measurement of behavior provides a less noisy signal of program success than the probabilistic impact a program may have on the target environmental outcome.

However, even when it may not be considered the final outcome, behavior is a critical intermediate output of any behavior change intervention. A proper PS-ToC, and therefore the indicators that follow from it, will not only include the final behavior that most closely causes the environmental outcome, it will also include each behavior identified as critical on the path to achieving that final behavioral goal. Once these behavioral indicators have been identified, they can be measured in a variety of ways (including direct observation, proxies, and self-reports), each offering unique benefits and costs. Please refer to the *Indicator Measurement* section for further discussion.

Psychological Indicators

Many program intervention elements aim to change behavior through the beliefs and preferences of their target actors. In traditional ToCs, these changes are often implicit, directly linking the intervention elements to the behavioral output. However, a PS-ToC makes explicit the intermediary step of belief and preference shift. By including this step, the PS-ToC acknowledges that even if the intervention element is delivered as directed, it may fail to change the necessary beliefs or preferences, and, therefore, fail to change behavior. The inclusion of psychological indicators tied to target beliefs and preferences is thus critical for understanding the degree to which a program was successfully implemented, and to identify adaptive management decisions that can or could be made.

Psychological Indicators in Farmer Training

An extension worker-led farmer training attempts to change three key target beliefs: (1) that overwatering is decreasing yields, (2) that watering three days a week is sufficient, and (3) that others in the community are watering less. All three of these beliefs are hypothesized to be the reasons that this training may lead participating farmers to waterless. They are, therefore, part of the PS-ToC, and psychological indicators to be incorporated into the monitoring and evaluation of that intervention component.

Social Indicators

Social indicators can be relevant at three distinct parts of a program's theory of change. First, social indicators can be key outcomes of a behavior change program. These can include what are commonly called objective outcomes, like how by restoring coastal habitats a program may aim to increase the livelihoods of those who rely on fisheries, as well as subjective outcomes, such as the program's effect on improving fishing communities' well-being. Second, a program may see changing social structure as both causing and being caused by behavior change—things like increasing trust in a particular institution or modifying the structure of a social network. Since changes

in both of these social indicator categories are critical for establishing why a program works and whether it has achieved its intended outcomes, it is critical for these social elements to also be captured in a PS-ToC and to serve as useful indicators for monitoring and evaluation.

Finally, social structures may represent the context in which an intervention takes place, even if this structure, in of itself, is not changed by said intervention. In other words, a particular social context may be critical for an intervention to work, but it might not itself be directly intervened on. This category is different from the two above, as those were expected to change as a result of the intervention. It can be important to measure these social contextual factors, as they can provide critical predictive information about whether an intervention is expected to work in a given context. For this reason, they should be represented in a PS-ToC. However, because they are not hypothesized to change over time, they should not be included as indicators for measuring program effectiveness or success.

Social Indicators in Farmer Interaction

A multi-community gathering of innovators is designed to change the social network farmers rely on for information on new environmentally-friendly farming techniques. The structure of target farmers' social networks, in relation to whom they rely on for information on farming techniques, is an integral part of the PS-ToC for the program. Changes to these social networks should be incorporated as an indicator of success relative said farmer-gathering program component. It is important to recognize that if the change to the social network is successful in changing behavior, that may result in a feedback loop further cementing that novel social network, a critical element to capture in the PS-ToC.

The same program may rely on trust in agricultural extension agents in order to convene the gathering. This is a key enabling social context that must exist for the program to succeed - it should be represented in the PS-ToC. However, because the program does not aim to change the degree to which the community trusts agricultural extension agents, it does not need to be monitored or evaluated for change.

Just as our understanding of behavioral and social science has advanced our understanding of the effective intervention design, so has it advanced our understanding of how we ought to monitor and evaluate them. Just as traditional ToCs forced program designers to be explicit about their previously implicit hypotheses, the PS-ToC forces that same explicitness with regard to changes in psychological and social factors. These explicit statements allow for the development of program-specific psychological and social indicators, which can then be incorporated into an understanding of the why of a program's success, as well as into rapid adaptive management efforts.

Figure 3 graphically depicts a PS-ToC example. It represents a component of the overall PS-ToC of Fish Forever, a global fisheries program that aims to increase fish stocks and community livelihoods in key coastal ecosystems for ten countries. The PS-ToC includes a mapping of various activities, as well as predicted changes to target actors' psychological and social states (these, in turn, are tied to key behaviors). By explicitly representing the intermediary social and psychological states a community is expected to experience, the PS-ToC provides clear guidance on the indicators that should be measured to assess whether a program is working, to diagnose how it might work better, and to create generalizable knowledge for future programming. The PS-ToC also includes explicit statements about the assumed enabling social conditions, which can be used in determining the intervention's applicability to a novel context.

It is important to recognize the bi-directional flow between behaviors and psycho-social states. While changes in psycho-social states cause changes in behavior, this process is dynamic, with those changes in behavior feeding back into the psychosocial context. This can be particularly pronounced and critical to capture in programs targeting social norm change, which is an inherently dynamic process whereby the changes in a segment's behavior results in changes in the wider social context.

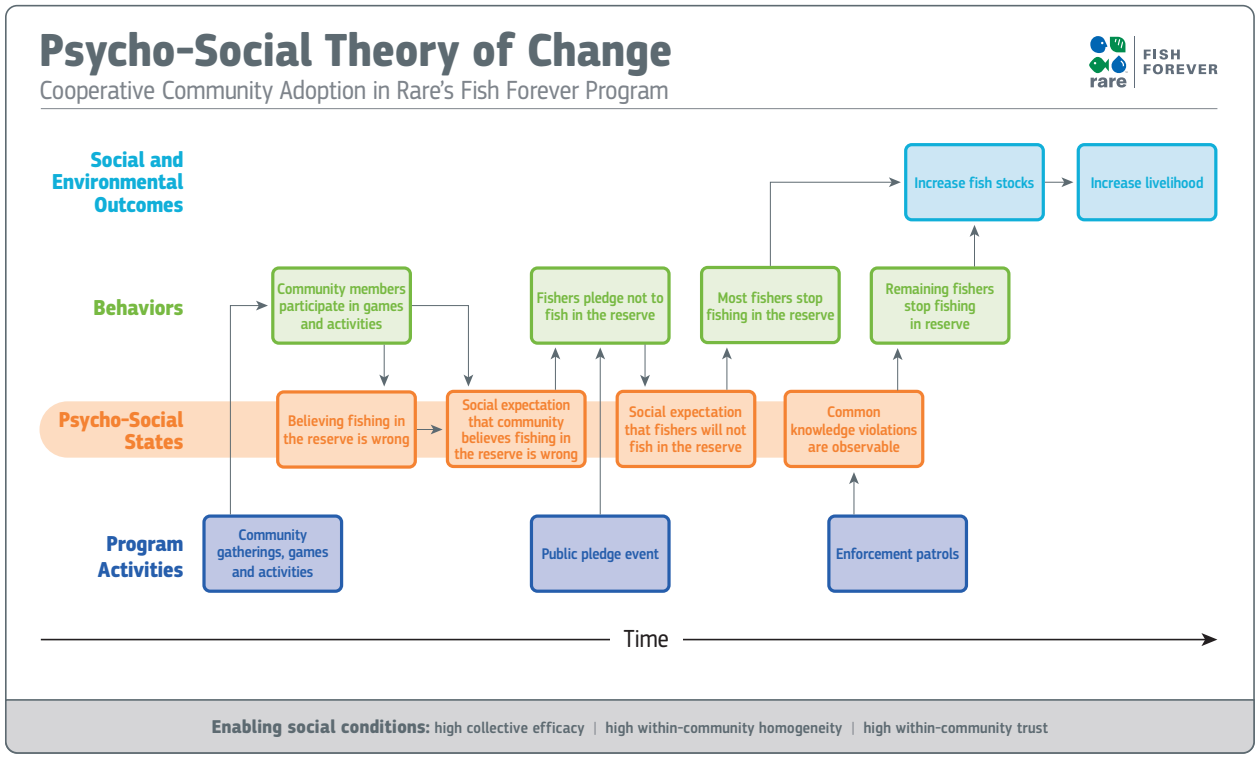

Figure 3: The psycho-social theory of change for Rare's Fish Forever program defines how program activities affect psycho-social states and behaviors, which in turn drive social and environmental outcomes. It also explicitly defines the enabling social conditions needed for the program to operate successfully.

Indicator Measurement

After establishing what indicators to measure, these need to be operationalized into assessable measures. Psychological indicators require introspective access, meaning we have to measure what is going on inside someone's head. This means that psychological indicators generally require self-report for measurement. This can be contrasted with behavioral indicators, which can be assessed with a variety of methods, including direct observation, where the behavior is directly measured; proxy measures, which assess outcomes which are assumed to be tightly related to the target behavior; and self-report measures, where the rate or intensity of a behavior is inferred through responses on instruments such as surveys. In this section, we review the benefits and drawbacks of each of these techniques and provide best practices to mitigate limitations.

Direct Observation

Direct observation represents the most straight-forward approach to measuring a behavioral indicator. However, direct observation is often limited by the observability of the behavior we want to monitor. Often, the behavior that program designers are interested in is the actual presence (or absence) of an illicit or private behavior. Such contexts require more creative, specialized design for successful direct observation.

As an example, take the use of toilets, a practice that has significant public health and environmental benefits. Toilet use has generally been seen as difficult—if not impossible—to measure through direct observation due to its inherently private nature. However, the introduction of PLUM devices, which use passive infrared motion detection to estimate toilet use rates, has provided a novel opportunity to directly evaluate toilet use behavior (Clasen et al., 2012). Used in this way, creative technological solutions can allow for the direct observation of practices previously seen as hard to assess.

Still, this does not mean that direct observation is without bias. Unless relying entirely on existing data sources collected without target actors' awareness, direct observation is subject to what is known as the Hawthorne Effect: where people's behavior is affected by the simple fact that they know they are being watched (Landsberger, 1958). Expanding on the previous example, the fact that PLUM devices allow for accurate observation does not address the problem that one might choose to behave differently depending on whether or not a PLUM device is installed on their toilet.

While direct observation is often seen as impossible for behaviors that are illicit, this is not always the case (given recent advances in remote sensing in particular; Kamminga et al., 2018). For example, the Eyes on the Seas initiative uses live satellite data to detect intrusion into marine protected areas in real-time (Kalinauckas, 2015). Technology has also made terrestrial poaching far more directly observable. For example, flying remote-operated craft over reserves in South Africa allows for the discrete and direct observation of rhinoceros poaching activity (Mulero-Pázmány et al., 2014). However, this style of direct observation is not always possible, and when it is, it may be prohibitively expensive. In these cases, proxy measures are reasonably employed to improve indicator measurement.

Proxy Measures

Proxy measures are those that, instead of directly measuring the behavior itself, measure some causal consequence. The prevalence of the behavior itself is then inferred from that consequence. These methods are most appropriate when direct observation is prohibitively costly or difficult to employ, and when the causal connection between the proxy and the indicator stands up to scrutiny.

An example of a successful proxy measure is assessing the amalgamation of various energy-saving behaviors through energy use, measured at the household meter (ex. Allcott & Rogers, 2014). A kilowatt-hour reduction is not

itself a behavior; instead, it is the consequence of a suite of behaviors that range from simple things like keeping the lights off all the way to replacing one's energy-inefficient appliances. Nevertheless, when an intervention targets these behaviors broadly, the measurement of household-level energy usage (relative to a valid control) represents a strong proxy measure for those behaviors—despite the measurement itself not being a direct measure of any of them. Many proxy measures have the added benefit that they are directly observable from already-collected data. Not only is this cheaper, it also means that any effect observed is unlikely to be the consequence of a Hawthorne style effect, where people are responding to having been unnaturally measured. Looking at an example from another domain, waste audits represent a similar tightly-linked behavior-proxy, making them particularly valid as proxy measures of consumption (Hoover, 2017). Waste audits generally involve examining the waste content from a household or institution, measuring the presence of different forms of waste. This measurement of waste (relative to a valid control) can be used to assess the relative effectiveness of waste reduction interventions despite never having to observe actual waste-reducing behaviors.

Some proxy measures are more dubious, however; often because of the role that other variables can play in influencing both the target behavior and any bias in proxy measurements. This is commonly observed in the use of enforcement logs, such as when arrest records are used to assess the prevalence of an illicit behavior. Imagine that an increase in enforcement incidents, like arrests, is observed. One cause may be a true increase in the prevalence of the illicit behavior. Another might be an increase in the presence of enforcement (which is, paradoxically, often the intervention itself). This means that one could observe an increase in enforcement actions because of an increase in the presence of enforcement, all while that same presence actually decreases the prevalence of the illicit behavior.

As demonstrated in these examples, proxy measures should be evaluated based on the degree to which they can be expected to result uniquely from the behavior of interest—and should be considered suspect if that is not the case. Additionally, the energy and waste examples, despite being strong cases, raise another common issue found in these measures: aggregation. While energy readings tell us that the household has adopted behaviors to reduce their energy consumption, they aggregate over a series of behaviors such that it is difficult to determine what unique actions were adopted among which members. Similarly, waste audits aggregate behavior across individuals and across time. This allows implementers to infer the average change in a behavior, but provides but a minimal understanding of how that change is distributed across individuals and situations.

Self-Report

Technically a form of proxy measurement, self-report measures ask the actor themselves to report on the prevalence of their behavior. These measures are subject to both random noise and bias due to cognitive and social factors. However, for many indicators, self-report may be the only practically available method. Additionally, many psychological and social indicators, by definition, require introspective access, meaning they must be assessed through self-report. Thankfully, a series of best practices can improve self-report measures by limiting, while not completely removing, the influence of these cognitive and social biases.

In Figure 4, Sudman et al. present a cognitive model of survey response, which helpfully illustrates the various opportunities for cognitive and social factors to add noise and bias to self-report measurement (1996).

Figure 4: Cognitive Model of Survey Response from Sudman et al. (1996)

The first four steps are particularly subject to cognitive bias, whereas response editing is primarily socially influenced. Designing valid self-report measures requires minimizing bias at each step in the process.

Example: Bias in Self-Report measures

We can explore each of these cases for bias introduction through the use of an example. Imagine we are attempting to measure the size of a fisher's average catch by asking them: "In a typical day, how many kilograms of fish do you normally catch?"

After being asked the question, the fisher must **interpret the language used**. For example, a fisher might interpret "a typical day" to mean only those days when the water is calm enough to fish, every day but Sunday, every day including weekends, etc. After interpreting the question, the fisher must then **retrieve the information from memory**. This is often done using cognitive shortcuts, which can bias their response. For example, a fisher might think back to a particularly salient day when they had a particularly large catch. They could alternatively think of yesterday's catch. Or perhaps vaguely think about how much a fisher would normally catch, rather than even thinking of their own one. To **translate that information into a judgment**, they must then conduct an internal calculation, perhaps by mentally counting the number of fish buckets they have on their boat and then estimating what proportion of those buckets was full on their last trip. They might also do this in a completely different way. And after that, they still have to **format their responses** to the response format that is made available to them. For example, if the question asks for an answer in kilograms, the respondent will need to estimate how many kilograms one bucket of fish represents and multiply that out. As has hopefully become clear, answering what might have looked like a simple question requires an extraordinary amount of mental effort—and every step risks further biasing one's answer. In order to improve the self-report process, items must, therefore, be made simple and unambiguous. While this advice may seem obvious on its face, meeting a sufficiently high degree of simplicity and clarity requires that each question be evaluated at each of the first four levels illustrated above: interpretation, retrieval, judgment, and format.

Unfortunately, simplicity and clarity are insufficient in dealing with the final question-answer step: **response editing**. Generally, response editing is the result of what we call the social desirability bias—that is, respondents are more likely to give answers that they believe are consistent with the surveyors' (and, more generally, society's) expectation about how they ought to have behaved. This bias is particularly pernicious given that it can be difficult to assess the magnitude of its effect. Various best practices and techniques do exist, however, that can minimize the bias' effect on participants' responses.

Minimizing social desirability bias starts long before the question is asked. The enumerator should ideally be from a neutral third party, rather than the implementing organization or the specific community of the respondent. While building rapport with the respondent, the enumerator should describe the goal of the data collection as being to better understand the participant, their beliefs, and motivations, rather than it being linked to an evaluation of the program. In the context of a particular self-report item, presenting both options as valid responses allows a respondent to feel less aberrant reporting their true behavior or beliefs. For example, one might add the preamble

"some people fish inside the MPA, and some people do not" before asking if the respondent does in order to validate both as reasonable answers. Additionally, answers can be made private, which minimizes the immediate uncomfortable nature of giving a socially undesirable response. This can be done in a variety of ways, such as having a respondent mark their answers on paper rather than verbally, or simply turning the enumerator's tablet towards the respondent so that the enumerator cannot see it (yet allowing the respondent to give their report directly).

While these techniques may minimize bias through validation and privacy, they do not provide anonymity, which is often felt to be necessary for extremely socially sensitive questions (e.g., those relating to illicit behavior). Anonymity can be provided through anonymous polling. However, anonymous polling only allows inference at the aggregate level, both reducing statistical power and eliminating the ability to conduct individual-level analysis. Environmental social scientists have applied a number of advanced survey techniques, including randomized response and unmatched counts (Dalton et al., 1994; Warner, 1965), to assess the prevalence of illicit behavior all while preserving their ability to conduct individual-level inference (e.g., Bergseth et al., 2017; Solomon et al., 2007). Additionally, though offering incentives for truthful responses cannot eliminate the pull of social desirability, they can—at the very least—offset its effects. Such incentives could be utilized, for example, when the direct monitoring of all respondents' behavior would be infeasible, but when occasional spot checks would otherwise be possible. In this case, an incentive can be tied to accurate reporting relative to the possible spot-check.

While self-report suffers from various possible threats to accuracy and precision, it may be the only possible option for logistical (such as cost) or logical (such as the indicator being psychological) reasons. In those cases, considering the various cognitive and social threats to validity outlined above, and designing to account for them, can improve the quality of self-report measures.

Framework for Indicator Monitoring and Adaptive Management

While monitoring and adaptive management remain frequent buzzwords for best practices, monitoring without clear intentions of how those data will be used for decision making undermines the overall effectiveness of an implementing organization. Similarly, some implementers collect monitoring data purely to satisfy the requirements of funding agencies, which represents a similarly ineffective—and unethical, as some have argued—approach (Gugerty & Karlan, 2018). However, the inclusion of psycho-social indicators derived from a program's PS-ToC presents new opportunities for effective use of monitoring data in two ways. First, the data can be used for refined program diagnostics. Second, it can be used to inform live programmatic decisions that present an opportunity for net-reduction in overall program costs.

Psycho-Social Diagnostics

Traditional monitoring for adaptive management generally tracks the degree to which a program was delivered according to plan and may also include the tracking of intermediate behavioral outcomes. However, this tracking is insufficient for programmatic diagnostics and adaptive design. This is because it is missing the critical monitoring of the intermediate psycho-social indicators linking intervention components to behavioral outputs. In these traditional monitoring approaches, when the data indicate that the program elements were implemented according to plan, but the intermediate behavioral output was not achieved, this monitoring provides no insight into why the element failed to result in the intended behavioral outcome. It could be because the intervention element failed to change the targeted psychological or social state, or it could be because the change in that psychological or social state was insufficient for behavior change. This difference is diagnosis is critical for continued intervention development.

An alternative is the frequent pulse monitoring of the community on key psychological and social indicators. By combining these data with the traditional monitoring of program delivery and intermediate behavioral outcomes, a program designer can then assess where it is in the causal chain that breakdowns may be occurring (e.g., whether they are between the implementation and the psycho-social indicator, or between the psycho-social indicator and the intermediate behavioral output). This data can then be used to revise programming, either through the improvement of program elements to increase their ability to have the intended psycho-social effect, or through a shift in the theoretical understanding of the problem to focus on other psycho-social targets.

Live Programmatic Decision Making with Dynamic Programming

Monitoring data is most frequently used in a retrospective capacity to improve delivery in future instances. However, it is also used to ensure present programs are delivered as expected. This style of rapid management often focuses narrowly on the elements of the program that are directly delivered by the implementer, such as whether particular training sessions took place or particular tools were distributed. This corresponds to rapid response to program delivery indicators traditionally found in a program's ToC.

The introduction of a PS-ToC presents an opportunity to conduct rapid response to psycho-social indicators to improve program efficacy and cost-effectiveness, a practice we refer to as dynamic programming. In a dynamic programming implementation, key psycho-social indicators of a program are assessed on a frequent basis throughout program delivery, a practice known as pulse monitoring. After an element of a program is implemented, the next round of pulse monitoring will indicate whether that program element has achieved the intended change in the psycho-social indicator. Pulse monitoring can also be used to assess feedback changes in the opposite direction in which changes in behavior cause changes in psychosocial states as hypothesized in the PS-ToC.

Tracking changes to the psycho-social landscape through pulse monitoring allows for rapid implementation decisions that can improve the effectiveness and cost of program delivery. While traditional methods of program delivery assume that for all units of intervention, whether they be individuals, communities, or geographies, the same dosage of an intervention is required to drive the needed change to psychological and social states. The dynamic programming approach acknowledges that there is heterogeneity in the dosage that an intervention requires.

This pulse tracking allows for two types of dynamic programming responses. If a unit is particularly responsive to one component of an intervention, the dosage of that component (for example, number of training sessions, the length of an incentive, etc.) can be reduced, as it has already achieved its desired effect. This has obvious cost-saving advantages. If a unit is particularly unresponsive to the standard dose of an intervention, either that dosage can be increased, or the program implementer can re-evaluate whether that particular component was actually appropriate for driving change in the target psychological or social state and whether expected feedback loops between behavior and psychosocial states are in fact present. While at first glance this may appear to be an increase in cost, but what would be even more costly is proceeding to the next step of an intervention without laying a proper foundation. Dynamic programming allows for the assurance that said foundation exists, thereby increasing program efficacy.

Methods for Pulse Monitoring of Psycho-Social Indicators

The pulse monitoring required for both improved psycho-social diagnostics and dynamic programming presents an additional cost to an implementing program. It is therefore important to identify methods that are scalable and cost-effective to be applied across programs, geographies, and social groups.

While many programs conduct in-person, pre-post interviews of program participants that can assess change in psycho-social indicators, these do not provide information useful for rapid programmatic responses. It would also be prohibitively costly for most programs to conduct frequent in-person interviews that gauge changes on these dimensions. We, therefore, present a scalable, technologically-assisted alternative.

Over the past decade, mobile phone ownership has skyrocketed across low-income countries. In 2019, Pew reported that the median emerging-economy country has almost 80% mobile phone penetration among adults. While remaining differential access to these phones based on a variety of social factors including income, gender, and social class needs to be accounted for, this high prevalence makes mobile phones an ideal candidate for scalable psycho-social indicator monitoring. This is in contrast to the prevalence of smartphones in emerging economies, which stands at only 45%. This discrepancy means that any measurement aiming for representative coverage needs to be administered over voice or text message, rather than through smartphone apps.

Automated text message surveying and interactive voice response (IVR) allow for the inexpensive large-scale monitoring of psycho-social indicators within a population. Researchers working in low-income contexts have found that both methods can be effective for high-frequency data collection, but that IVR tends to result in higher response rates as compared to text messages (Ballivian et al., 2015). By collecting these psycho-social measures on a regular basis over the life of the program, the implementer can track their changes live over time, and make informed rapid program delivery decisions.

The Development of a Dynamic Programming Framework for Coastal Fisheries Management

Rare's Fish Forever program described earlier is presently integrating pulse monitoring for dynamic programming into its 10-country program. As each site in the program presents unique programmatic needs, it presents an excellent opportunity to optimize differing dosages and delivery modes for the program's various components.

The Fish Forever community engagement program can be divided into three stages outlined in the Theory of Cooperative Behavior Adoption: generating collective demand, coordinating a shift in behavior, and strengthening the social norm (Thulin, 2020). This resulted in the following key psycho-social indicators in the Fish Forever PS-ToC for the key behavior of only fishing outside the reserve area.

Generating Collective Demand:

- Psychological: Believing that it is wrong to fish in the reserve
- Social: Common knowledge that the large majority of community members believe it is wrong to fish in the reserve

Coordinating a Shift in Behavior:

- Social: Common knowledge that all fishers in the community will no longer be fishing in the reserve
- Behavioral: Fishers are only fishing outside the reserve

Strengthening the Norm

- Psychological: Believing that if one were to fish in the reserve, the rest of the community would find out
- Social: Common knowledge that all members are expected to stay outside the reserve, and that intrusion would be socially sanctioned

It is critical for the PS-ToC logic of the Fish Forever program that these three steps occur in sequence. It is therefore important that the psychological and social indicator thresholds be met in the '*Generating Collective Demand*' step before proceeding to the '*Coordinating a Shift in Behavior*' step. Given the incredible diversity of sites in which the Fish Forever program is deployed, implementers cannot rely on the same timeline for progress on these steps in every site.

To address this challenge, Rare partnered with the telecommunications service company EngageSpark to develop an IVR based solution that would measure each of these six indicators on a bi-weekly basis among a panel of respondents within each community in the Fish Forever program. This bi-weekly feedback is designed to directly feed into the program delivery decisions of the implementer, telling them when it is the appropriate time to proceed to the next step in the cooperative behavior adoption process. When a community does not experience the expected changes in target indicators, a standardized qualitative methodology is employed to determine why the programming is not working. This may result in either a change in the method of delivery,

or simply, in additional time spent on those components of the intervention specific to that community until targets are achieved.

The Fish Forever dynamic programming platform is currently undergoing testing, with the goal of program-wide rollout in 2021.

Figure 5 graphically depicts how pulse monitoring can be incorporated into live program delivery decisions, using Rare's Fish Forever program as an illustrative example. Throughout each phase of the program, the pulse monitoring assesses the degree to which target psychosocial states have been achieved. If they have not, and the site is therefore behind schedule, the implementation team then conducts a rapid qualitative assessment to determine why the intended psychosocial states have not been achieved. This assessment is then incorporated into localized revisions to the program activities, which are then executed until the target state is reached.

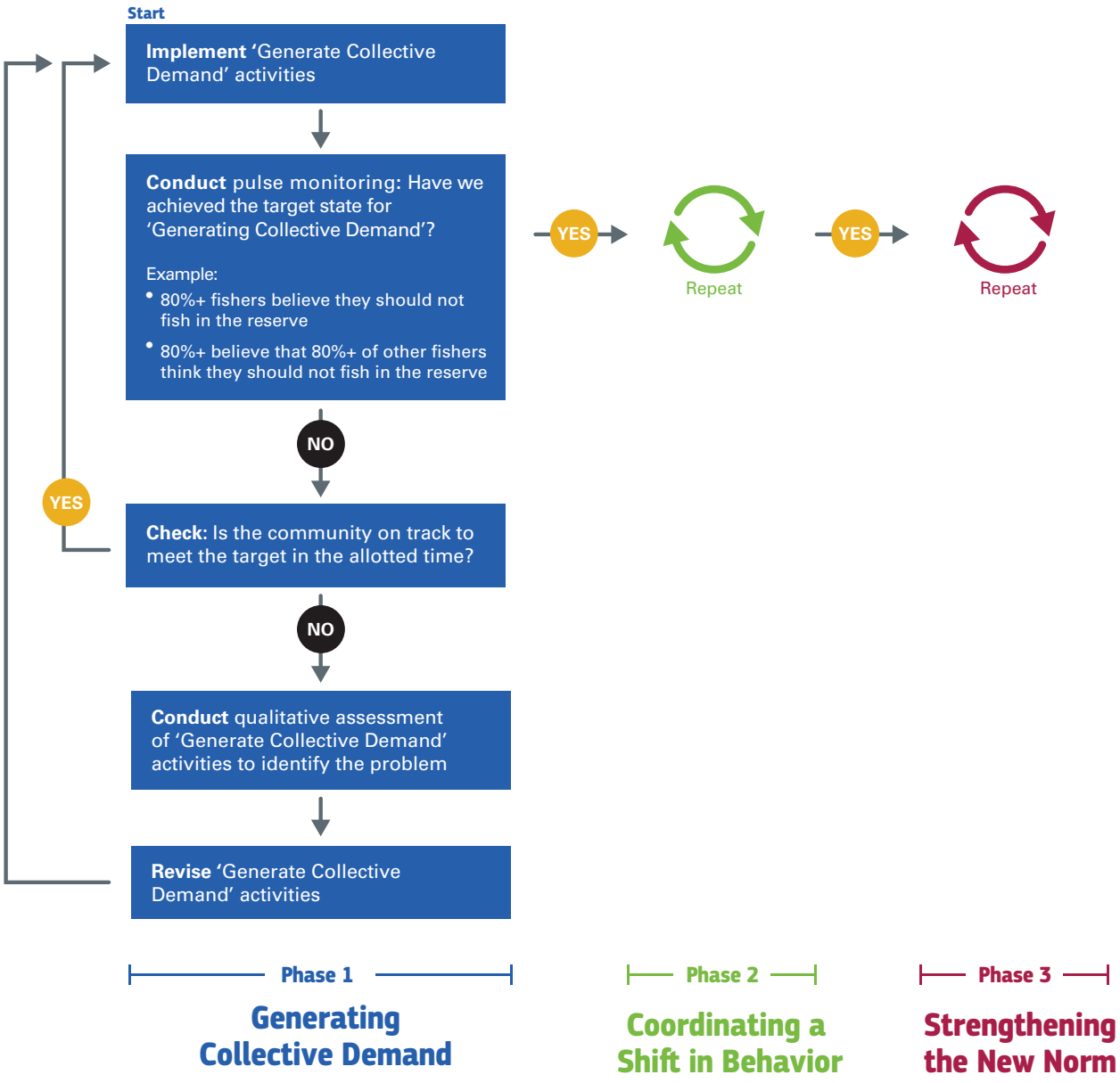

Figure 5: Flowchart representing the steps in the Dynamic Programming process of Rare's Fish Forever program. Dynamic Programming represents the inclusion of pulse monitoring, where the psychological and social state of the community is continually tracked over time to improve live programmatic decision making such as when to advance to the next phase in a program.

Frameworks for Evaluating Changes in Indicators

Evaluation serves a critical role in a behavioral intervention lifecycle. After the development and testing of a program, rigorous evaluation allows for strong causal claims about the relative effectiveness of the intervention, as compared to either a control or a common "standard of care" intervention. This is an important step for a number of reasons. First, without robust estimates of the treatment effect of an intervention, it is impossible to calculate the cost-effectiveness of a program. Second, it is critical from an ethical perspective: properly determining that a program is causing an observed change is a necessary input into whether the scaling up of that program is ethical in terms of both the deployment of resources as well as the way we treat target communities.

It is important to acknowledge that this is not to say that every deployment of every intervention needs to be evaluated against a counterfactual. Similar to how many medicines have well enough documented effects that we no longer need to assess their effectiveness, some behavior change programs may be similar. The key question to answer is whether a particular program has been demonstrated in a sufficiently similar context to confidently estimate the size of its effect without any evaluation. The most common situation in which this is the case is when a program has been demonstrated as successful for a random subset of a target population, and then, is deployed for that entire population. It is almost never the case when a program is applied to a new social or ecological context that it will be as successful as before.

There are various designs for causally inferring the treatment effect of an intervention. Each of these methods is fundamentally attempting to ask the question: how would those in the population have responded to a particular intervention compared to a counterfactual world where no treatment was given at all? This difference tells designers what the causal effect of that intervention is. In this section, we aim to aid in the initial evaluation of these options for a given program by giving an overview of their logic as well as their relative costs and benefits to other possible evaluation designs.

Randomization Based Methods

Randomization is commonly seen as the gold standard in causal inference evaluation as it relies on the fewest assumptions for unbiased estimates. We therefore present randomization-based methods first, as all other methods of casual inference will be compared to them in terms of what additional assumptions they require.

In this section we review Randomized Controlled Trials (RCTs) and Clustered Randomized Trials (CRTs). RCTs are normally used when an intervention can be randomized at the individual or household level. CRTs are used when interventions can only be randomized at some group level, such as with a community or region, but the effect of that intervention can still be measured at the individual or household level.

Randomized Controlled Trials

Fundamentally, RCTs require units, often individuals, to be randomly assigned to a particular treatment condition. These can be the intervention of interest, a standard control condition (meaning no treatment), or some other treatment to which the treatment of interest is compared (this is sometimes referred to as an active control). To evaluate the effectiveness of an intervention, the average value of the control group's indicator of interest is subtracted from the average value on the indicator in the treatment group. The remaining value is the average treatment effect, representing the change in the indicator that can be causally attributed to the treatment.

The strength of RCTs rests on a fundamental assumption of randomization. At its core, this assumption is that every unit in the population of interest has an equal chance of being selected to participate in the study, and every unit in the population that has been selected has an equal chance of being assigned to each treatment condition. This means that if a partner organization only wishes to work in particular communities, then any RCT run within that setting is only technically generalizable to those particular communities rather than to the wider population. It also means that if any preference is given for any units to be assigned to the treatment versus the control condition, the validity of inferences is undermined.

Method for Cost Savings in RCTs: Unequal Group Sizes

Those who are designing evaluation protocols often assume that the most cost-efficient way to do so is to allocate the same number of units to the treatment and control conditions. However, this is not always the case. Specifically, this does not hold true when the cost of recruiting and evaluating an intervention unit is different from the cost of a control unit. For example, an intervention evaluating the effectiveness of delivering clean cookstoves may find that the cost of delivering, training, and surveying in the treatment condition is $90 per participant, whereas the cost of only surveying in the control condition is $10. In this case, the most cost-efficient allocation is to recruit three times more respondents in the control condition than in the treatment condition. The method for calculating this optimal allocation can be found in List et al. (2011).

Clustered Randomized Trials (CRTs)

Discussion of RCTs, like the above, generally focuses on their application to programs that target individuals, and critically, that can be randomized at that individual level. This is highly applicable for those behavior change programs that are suitable for individual or household level randomization. However, due to their fundamentally geographic and often community-focused nature, many environmental programs are not suitable to be delivered at that level of granularity, and instead, must be administered at the community or regional level. CRTs present a randomized design framework suitable for assessing such interventions.

The setup of a CRT is quite similar to an RCT: clusters, such as communities, are randomly assigned to either receive the intervention of interest or some comparison control. The key difference with RCTs rests on randomization at the cluster rather than individual level. And this has significant implications for the study design: depending on the degree of similarity between individuals within each cluster, CRTs can require a much large number of clusters. Simply having two communities with individuals from one community assigned to one treatment and individuals in one other community assigned to another treatment, even with a high number of individuals monitored within each community, will never be sufficient for valid statistical inference.

Thankfully, if a sufficient number of clusters are recruited to participate, CRTs present causal inferences that are as strong as those of RCTs—at least for estimating the average treatment effect of an intervention given the minimal assumptions tied to randomization.

Quasi-Experimental Methods

While randomization-based experimental methods offer strong causal inference with minimal assumptions, randomization is not always an available evaluation strategy. This has led to the development of quasi-experimental

strategies that attempt to infer the causal effect of an intervention without randomization, albeit with additional assumptions that are generally impossible to assess directly.

Difference-in-Difference

Difference-in-difference is a commonly employed strategy when the assignment to a condition cannot be randomized, but when baseline data is available for both those who receive the treatment as well as for a group that was not assigned to receive the treatment but that appears otherwise similar. Instead of simply comparing outcome indicators after the treatment has gone into effect, difference-in-difference compares the change in indicator measures of the treatment group with the changes in that of the control group. The average treatment effect of the intervention is then estimated as the change in an indicator of the control condition value subtracted from the change in the indicator's value in the intervention condition.

This comparison of differences, rather than just raw outcomes, is critical for the logic of difference-in-difference. Because of this focus on change, difference-in-difference does not need to assume that groups assigned to the intervention and those assigned to the control are at a similar level on the outcome indicator. Instead, the most critical assumption of difference-in-difference is parallel trends. This assumes that, in the absence of any intervention, those in the intervention group would experience the same amount of change in outcome indicators as those in the control group. This assumption is generally not testable, however, and must be argued on a case by case basis.

Matching

Matching techniques are an additional method for overcoming the hurdle of not being able to randomize units to a control and intervention condition. Instead of randomizing, matching attempts to build a control condition by identifying units that are similar to each of the treatment units—doing so relative to a set of observable characteristics.

There are multiple methods for finding matches. Two common methods are exact matching and propensity-score matching (PSM). Exact matching relies on being able to identify units to use as the control group which show the exact same values on each of the observable characteristics (such as someone's gender, SES, distance from a resource) for the treatment units, except for the fact that they did not get the intervention. Alternatively, PSM focuses on those characteristics that are likely to predict receiving the intervention—matching people based on their likelihood of receiving the intervention rather than on the underlying characteristics that define it. This means that for each unit evaluated in the treatment condition, there is a unit in the control condition that, because of its baseline characteristics, would have been expected to be in the treatment condition (i.e., had a similar likelihood) but was not.

The drawback of matching methods is that they are only matched on observable characteristics, and these almost never include all of the characteristics that could explain why a unit received the treatment instead of being assigned to the control. This failure to include all (not just observable) explanatory variables means that any matched comparison could be subject to bias because of those unobservable characteristics yet unaccounted for.

Pre-post Comparison

Pre-post comparison involves estimating the effect of an intervention by simply subtracting the average baseline value of the outcome indicator from the average post-intervention value of the outcome indicator. While this technique is commonly used to evaluate environmental interventions, it rarely ever renders valid estimates.

Pre-post comparison makes the incredibly strong assumption that no other factor influenced the value of the outcome indicator between the pre and post measurements except the intervention itself. This assumption is rarely ever true, and even more rarely ever justifiable.

One case in which this may hold for a behavioral intervention is when a new practice is being introduced, and no other actors are operating in the area which might introduce that practice. This could be seen, for example, in the adoption of a novel product, where that product is available from those implementing the intervention and no other sources. However, this narrow case rarely represents the practices that environmental behavioral interventions seek to address.

Additional Methods

While the randomized and quasi-experimental methods described above represent the large majority of techniques used to evaluate environmental programs, they are not exhaustive. Additional econometric techniques, which are often used to evaluate policy interventions, are also sometimes applicable, including regression discontinuity, instrumental variable analysis, and synthetic controls. For those wishing to dive further into these techniques, Cunningham (2021) provides an excellent introduction.

Disaggregating Impact Across Social Differences

Due to behavior change programs targeting specific barriers and motivations, and people with different social identities having different barriers and motivations, behavior change programs can have dramatically different effects for different social groups. Traditional methods of evaluation simply aggregate the entire sample across these groups together, returning the average treatment effect of an intervention. However, this average hides the nuance of whom is being affected by an intervention. It is therefore critical to analyze and report disaggregated estimates of program effects, which show how an intervention may have impacted different social groups differently. A key factor in identifying on what dimensions an effect should be disaggregated is to return to the program's foundational analysis of the socio-ecological context to identify any groups which might experience different motivations and barriers due to their social position. This disaggregated analysis is critical for two reasons. First, it allows for far more nuanced learning to incorporate into adaptive management of a program to increase effectiveness. Second, it allows implementers to evaluate the equitability of program outcomes.

Cross-Context Generalizability

The methods above provide varying levels of strong causal inference. However, all of these methods are subject to the constraint that the estimated intervention effects they identify are only representative of the population from which they were drawn. This issue is known as the generalizability puzzle: if something works in one socio-ecological context, will it apply in another? Thankfully, recent advancement in this space helps us begin to answer that question. Too often, those considering the scaling of a program focus on exclusively on geography, asking whether a particular intervention was tested in the same country or region. However, in order to answer the question of whether an intervention will work in a particular context, instead of asking where an intervention has worked, it is critical to ask why the intervention worked (Bates & Glennerster, 2017).

This returns to the earlier discussed Psycho-Social Theory of Change. While a traditional ToC gives very little insight into why an intervention worked, a PS-ToC explicitly states the why as a hypothesis. By doing so, one can then ask the question, does this why apply to this new context?

Bates and Glennerster lay out four steps for evaluating the generalizability of an intervention to a new context, adapted here:

1. Does the intervention have a Psycho-Social Theory of Change?
2. Given the PS-ToC, do the enabling conditions, including socio-ecological conditions, hold?
3. Given the PS-ToC, are you confident that the underlying psycho-social logic linking intervention components to behavioral outputs applies to the members of the new target population?
4. What is the evidence that the implementation process can be carried out with high fidelity?

While traditional evaluation will not tell designers whether an intervention will generalize, this four-step process integrates evaluation with our mechanistic understanding of the social and psychological context to allow for a more generalized understanding.

Conclusion: Recommendations on Indicators For Behavior Change Programming

We conclude with a synthesis of the content above into a series of best practices for the development, monitoring, and evaluation of indicators for behavior change programming.

1. In addition to the traditional components of a theory of change, the development of behavior change programming should include the adoption of a psycho-social theory of change (PS-ToC). The PS-ToC should include explicit representation of the psychological and social changes expected as the result of each element of a program between that element of the program and the behavioral output, as well as the necessary enabling conditions.

2. Program indicators for behavior change programming should include indicators for psychological and social elements of the PS-ToC linking intervention components to behavioral outputs.

3. Behavioral indicators are ideally measured through direct, unobtrusive observation. However, when direct observation is infeasible, carefully selected proxy measures of behavior and carefully drafted self-report measures may also be valid. Proxy measures should be closely, logically linked to the target behavior, and survey measures should be administered and drafted in a manner that accounts for best practice in addressing cognitive and social response biases.

4. Measures of psychological and social indicators often require actors to reflect on their own mental states and subjective view of their social circumstances that influence their behavior, and therefore can often only be measured through survey items. These, too, should be drafted applying best practices to reduce cognitive and socially biased responses.

5. Behavior change programming should adopt rapid, pulse-style, assessment of changes in psychological and social states of the communities in which they are deployed. This will allow for better adaptive management that improves future behavior change programming, as well as allowing for dynamic programming where an intervention can be designed to rapidly incorporate live psychological and social state information to improve a program's efficacy and cost-effectiveness.

6. Randomized evaluation, either through randomized controlled trials or clustered randomized trials, is the method of estimating program effectiveness that requires the fewest statistical assumptions, and should, therefore, be adopted for the evaluation of behavior change programs whenever possible. Quasi-experimental methods, such as difference-in-difference and matching, are also valid methods for inference but require additional, often untestable assumptions. Pre-post comparisons are rarely logically justified, and are very likely to yield statistically biased estimates. They should thus be avoided in the vast majority of cases.

7. Generalizability of a program across contexts cannot simply be derived from the statistical evaluation of a program. Nor can we assume that because a program has worked in one context, it will generalize to neighboring geographies. Instead, we must rely on our understanding of the enabling social and material environment as well as the program-specific psychological and social linkages between program intervention and behavioral output, which describes the why a program works. In order to evaluate whether a program will generalize to a particular novel context, it must be believed to be sufficiently similar for that same why to apply.

References

Allcott, H., & Rogers, T. (2014). The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. American Economic Review, 104(10), 3003–3037. https://doi.org/10.1257/aer.104.10.3003

Ballivian, A., Pedro Azevedo, J., & Durbin, W. (2015). Using Mobile Phones for High-Frequency Data Collection. In Mobile Research Methods: Opportunities and challenges of mobile research methodologies. Ubiquity Press. https://doi.org/10.5334/bar.c

Bates, M. A., & Glennerster, R. (2017). The Generalizability Puzzle. Stanford Social Innovation Review. https://ssir.org/articles/entry/the_generalizability_puzzle

Bergseth, B. J., Williamson, D. H., Russ, G. R., Sutton, S. G., & Cinner, J. E. (2017). A social–ecological approach to assessing and managing poaching by recreational fishers. Frontiers in Ecology and the Environment, 15(2), 67–73. https://doi.org/10.1002/fee.1457

Clasen, T., Fabini, D., Boisson, S., Taneja, J., Song, J., Aichinger, E., Bui, A., Dadashi, S., Schmidt, W.-P., Burt, Z., & Nelson, K. L. (2012). Making Sanitation Count: Developing and Testing a Device for Assessing Latrine Use in Low-Income Settings. Environmental Science & Technology, 46(6), 3295–3303. https://doi.org/10.1021/es2036702

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2010). A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009: American Journal of Evaluation. https://doi.org/10.1177/1098214010389321

Cunningham, S. (2021). Causal Inference: The Mixtape. Yale University Press.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the Unmatched Count Technique (uct) to Estimate Base Rates for Sensitive Behavior. Personnel Psychology, 47(4), 817–829. https://doi.org/10.1111/j.1744-6570.1994.tb01578.x

Gugerty, M. K., & Karlan, D. (2018). The Goldilocks Challenge: Right-Fit Evidence for the Social Sector (1 edition). Oxford University Press.

Hoover, D. (2017). Estimating Quantities and Types of Food Waste at the City Level. Natural Resources Defense Council.

Kalinauckas, A. (2015). Eyes on the seas. Engineering Technology, 10(3), 68–69. https://doi.org/10.1049/et.2015.0353

Kamminga, J., Ayele, E., Meratnia, N., & Havinga, P. (2018). Poaching Detection Technologies—A Survey. Sensors, 18(5), 1474. https://doi.org/10.3390/s18051474

Landsberger, H. A. (1958). Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. Experimental Economics, 14(4), 439. https://doi.org/10.1007/s10683-

011-9275-7

Mulero-Pázmány, M., Stolper, R., Essen, L. D. van, Negro, J. J., & Sassen, T. (2014). Remotely Piloted Aircraft Systems as a Rhinoceros Anti-Poaching Tool in Africa. PLOS ONE, 9(1), e83873. https://doi.org/10.1371/journal.pone.0083873

Solomon, J., Jacobson, S. K., Wald, K. D., & Gavin, M. (2007). Estimating Illegal Resource Use at a Ugandan Park with the Randomized Response Technique. Human Dimensions of Wildlife, 12(2), 75–88. https://doi.org/10.1080/10871200701195365

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology (pp. xiv, 304). Jossey-Bass.

Thulin, E. (2020). Cooperative Behavior Adoption Guide: Applying Behavior-Centered Design to Solve Cooperative Challenges. Rare.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60(309), 63–69. https://doi.org/10.1080/01621459.1965.10480775

Glossary

Actors: People whose behavior directly or indirectly affects program outcomes

Adaptive management: a process of updating and improving how a program is managed based on data and feedback about what is working or not working

Attitude: An evaluation of something, ranging from negative to positive

Barriers: Forces, attitudes, beliefs, or other reasons that prevent someone from doing a behavior

Behavior: An action that a person takes in response to something (a stimuli)

Behavior-actor pair: A grouping that identifies a behavior and who is doing it

Behavioral insights: The findings that result from analyzing patterns in how people tend to behave.

Behavioral system: A network of actors, mapping how each actor's behavior influences each other's capacity to act and their interest in doing so

Behavior change approach: A methodology for changing behavior, often drawing upon principles of behavioral design

Behavioral indicator: A measurement that signifies behavior (or something that approximates it) has changed

Behavior change intervention, programming: A set or sequence of activities that aim to influence actors to adopt target behaviors to achieve a certain outcome

Behavioral design: An approach that blends insights from the design and behavioral and social science fields

Belief: Something that someone accepts to be true

Bias/Cognitive bias: A way of thinking that systematically deviates from rational choice

Control: A level of an independent variable a person or group is assigned to in a study that receives no additional intervention

Counterfactual: A comparison for an intervention to assess its impact that shows what would have happened if no intervention had taken place

Cross-context generalizability: The degree to which something applies to other socio-ecological contexts

Design thinking: A creative and iterative process for developing, designing, and testing innovative solutions, often used in combination with human-centered design

Difference-in-difference: A quasi-experimental method that compares the pre-post change (difference) in outcomes for the treatment group with the change in outcomes of a comparison group

Direct observation: Type of behavior measurement based on directly observing behavior, rather than using a proxy or self-report measurement

Disaggregation: A data reporting process that shows how an intervention may have impacted different groups differently

Doer/non-doer analysis: A comparison of the motivations and barriers for people who are already doing the target behavior and those not doing the target behavior

Durability: The degree to which an intervention's effects persist during an intervention period and after the intervention has ended

Dynamic programming: Making live programmatic decisions about phase transitions, expansion, or termination based on real-time monitoring of psychological and social states of the target actors

Human-centered design: An approach or mindset to problem-solving that centers people's needs and goals in solution designs, often combined with design thinking

Matching: A quasi-experimental method that builds a comparison group by identifying units that are similar to each of the treatment units based on a relative set of observable characteristics

Motivations: Forces, attitudes, beliefs, or other reasons that encourage someone to do a behavior

Outcomes: The behavioral, social, or other goals or objectives a program is trying to achieve

Outputs: The components of a program that help to show how it achieved its outcomes and may serve as intermediary objectives

Pre-post comparison: A study where a treatment effect is estimated by subtracting the base-line value from the value after treatment

Program activities: The parts of an intervention that are implemented to change behavior, such as training sessions, pledges, incentive mechanisms, etc.

Prototype: A small-scale version of a behavioral solution that captures its essential features and can be tested with target actors

Proxy measures: Type of behavior measurement that uses outcomes assumed to be tightly related to the target behavior

Psychological indicator: A measurement that signifies a belief, attitude, or preference (or something that approximates it) has changed

Psycho-social state: Beliefs, values, expectations, and social relations that result from program activities, and other psycho-social states and also influence future behavior

Psycho-social theory of change: A theory of change that links intervention components to psychological or social changes, leading to behavioral outputs and environmental and social out-comes

Pulse monitoring: Assessing key psycho-social indicators on a frequent basis throughout pro-gram delivery

Quasi-experimental methods: Evaluation methods that infer the causal effect of an intervention without randomization when assigning individuals to treatment conditions

Randomized evaluations, Randomized Control Trials (RCTs): Evaluation methods where individuals are randomly assigned to treatment conditions

Self-report measures: Type of behavior measurement where the rate or intensity of a behavior is inferred through responses from instruments such as surveys

Social indicator: A measurement that signifies a social state, structure, or factor (or something that approximates it) has changed

Social marketing: The application of techniques from marketing to shift behavior to benefit individuals and society

Socio-ecological system: A system of interdependent linkages between ecological factors, social and cultural factors, and institutions at different scales that continually adapt over time

Stakeholders: Individuals or groups who have an interest in environmental outcomes or will be affected by a project and program

Study condition: A level of an independent variable a person or group is assigned to in a study

Study treatment: The intervention an individual or a group receives, based on the condition to which they were assigned

Systems thinking: An approach that synthesizes how parts of a system relate to, influence, and cause one another, often through feedback loops

CENTER FOR BEHAVIOR & THE ENVIRONMENT

Rare inspires change so people and nature thrive. Conservation ultimately comes down to people – their behaviors toward nature, their beliefs about its value, and their ability to protect it without sacrificing basic life needs. And so, conservationists must become as skilled in social change as in science; as committed to community-based solutions as national and international policymaking.

The Center for Behavior & the Environment at Rare is translating science into practice and leveraging the best behavioral insights and design thinking approaches to tackle some of the most challenging environmental issues. Through partnerships with leading academic and research institutions, they are bringing the research into the field to connect the next generation of behavioral scientists with practitioners on the front lines of our greatest environmental challenges.

To learn more, visit behavior.rare.org

STAP SCIENTIFIC AND TECHNICAL ADVISORY PANEL
An independent group of scientists that advises the Global Environment Facility

gef UN environment programme

The Global Environment Facility (GEF) was established on the eve of the 1992 Rio Earth Summit to help tackle our planet's most pressing environmental problems. Since then, the GEF has provided close to $20.5 billion in grants and mobilized an additional $112 billion in co-financing for more than 4,800 projects in 170 countries. Through its Small Grants Programme, the GEF has provided support to nearly 24,000 civil society and community initiatives in 133 countries.

The Scientific and Technical Advisory Panel (STAP) comprises seven expert advisers supported by a Secretariat, which are together responsible for connecting the GEF to the most up to date, authoritative, and globally representative science. The STAP Chair reports to every GEF Council meeting, briefing Council members on the Panel's work and emerging scientific and technical issues.