



Systems Evaluation Forum Synthesis of issues, discussions and resources

By Tanya Beer, associate director, Center for Evaluation Innovation

INTRODUCTION

In May 2017, the Gordon and Betty Moore Foundation convened a significant portion of its staff, together with a set of guest evaluators and systems thinking experts, for a two-day Systems Change Evaluation Forum. The aim of the Forum was to help establish a common understanding of systems change and its evaluation and explore a range of methods, tools, frameworks and approaches that could be useful to program staff. The Foundation's Measurement, Evaluation and Learning (MEL) team, along with the Center for Evaluation, also aimed to provide guest evaluators with an opportunity to network and share methods and tools for systems change planning and evaluation.

In many ways, the Forum was an experiment. We wanted to test what it takes to trigger deep dialogue among independent program staff who don't necessarily experience much commonality in their work or learn together in an intentional, coordinated way. We also wanted to test whether it was possible to seed rich learning between *evaluators* from different fields and philosophical orientations, and who are sometimes in competition and rarely have the opportunity to learn together. A objective was to lay the groundwork for a more open conversation within the foundation about people's challenges, doubts, and hopes for the foundation's approach to measurement and evaluation.

Some of these experiments bore out as we intended but others did not. The Foundation's MEL staff gained insights about important organizational tension points, learning needs, and effective learning formats through this experience. More importantly, the MEL team and meeting facilitators hope that the program staff came away with a clearer sense of what they need to do next (and NOT do!) with respect to learning together and improving their measurement and evaluation practices.

Day 1 focused on establishing a shared understanding of common systems change evaluation challenges and exploring how those challenges concretely play out in program staff's unique areas of work, using the foundation's Wild Salmon Ecosystem Initiative (WSEI) as one case example. Guest evaluators then led table conversations about tackling challenges identified during a series of one-on-one staff interviews earlier in the Spring. Resources provided by evaluators in those discussions are included in this paper. Day 2 offered a deeper dive into four dilemmas raised on the first day: causal analysis challenges, integrating learning and strategy, setting system boundaries, and finding meaningful interim outcomes in the face of long problem horizons. The Forum closed with Moore staff and evaluators separating out to discuss insights and lessons from the two days and what the foundation (or, in the case of the evaluators, the field of evaluation) needs to do next to improve its systems change evaluation efforts.



Rather than a traditional chronological recounting of proceedings, this paper synthesizes the core observations and dilemmas surfaced at the Forum about evaluating systems change work at Moore, offers resources to address some of these dilemmas, and suggests learning questions and decision points that might advance the foundation's cross-program learning and evaluation practice. It also includes links to resources related to each dilemma.

SYSTEMS CHANGE EVALUATION DILEMMAS AND SOLUTIONS

Through individual interviews and group meetings with staff prior to the Forum, we identified several common measurement and evaluation challenges that staff grapple with in their systems change efforts. We structured the Forum to address four of these in a variety of ways:

1. How do we better understand our unique contribution to change when there are so many other actors influencing the system?
2. How do we select and capture meaningful interim outcomes to understand systems change progress when we are operating in short time horizons and using strategies that have indirect impact?
3. How do we set realistic boundaries around the aspects of the system that our strategy and our evaluation should focus on?
4. How can we embed continuous, intentional learning into our strategy and evaluation work?

In preliminary interviews, staff also shared common internal operational and organizational challenges that sometimes stymie their measurement, evaluation and learning efforts:

- Having too little—or too much—up-front planning for evaluation and measurement at the outset of an initiative.
- Uncertainty about what kinds of evaluation questions will return actionable conclusions beyond confirming what program staff know intuitively already.
- Discomfort among some with the level of strategic flexibility that systems change efforts require, which in turn can force changes to the interim outcomes.
- Lack of nimbleness of the internal grants system to allow for easy, intuitive analysis of data *across* grants or at the portfolio level.
- Inconsistent approaches to documentation of findings and lessons.
- Uncertainty about how to integrate rigorous qualitative data and analysis with quantitative methods.
- Managing tension between what the board wants to see by way of outcomes and results and what program staff need to know for strategic and tactical decision making, given that staff have limited bandwidth and resources for evaluation and data collection.



Importantly, not all program areas experience all of these challenges because not all are engaged in systems change strategies. Even so, all program areas experience at least one or two of these challenges. For example, it may be straightforward to assess the unique contribution of Moore’s grants to individual investigators for science discovery, but still difficult demonstrate measurable impact within the timeframe of the grant; discoveries may not bear fruit for years to come. Program staff working to build cross-disciplinary collaboration in the quantum sciences may have to pay attention to a narrowly bounded system, but still be challenged to assess whether norms are really changing or the new way of working will disappear as soon as Moore’s funding does.

The Forum explored these dilemmas through a mix of case studies, small table talks about their manifestations in specific program areas, presentation of tools and approaches by evaluators, and some deeper dive, self-organized discussions. It was our hope that after the forum, staff at the foundation—including those who did not attend—will be able to adapt and apply some of these ideas to their own work. Below we synthesize four of these dilemmas, outlining the nature of the dilemma, examples of how it appears at Moore, solutions or tools offered at the forum, and key points of discussion and debate at the Forum.

Dilemma #1: How do we better understand our unique contribution to change when there are so many other actors influencing the system?

The nature of the dilemma

Understanding and isolating causality in complex systems can be enormously challenging. A traditional scientific paradigm that relies on experimental or quasi-experimental research design to establish causality is arguably ill-suited (if not impossible) for many types of systems change interventions, such as efforts to change public or institutional policy, market behaviors, or habits of institutional collaboration. It is difficult to create meaningful control groups or robust counterfactuals under conditions where many actors—including any who might fall into a control—are influenced by a wide range of forces in often unpredictable and invisible ways.

Additionally, the Foundation’s strategies are composed of many different interventions that are intended to work *together*—or to complement work supported by others in the field—rather than working in isolation to create change. For example, a policy change strategy may have only resulted in a “win” because it built on research and public education supported by other funders. It is possible that neither strategy alone could have produced the desired result.

So, while experimental and quasi-experimental designs can be useful to determine whether a more narrowly defined outcome can be attributed to an individual project (e.g., did patients receiving a particular approach to care at a particular hospital have better health outcomes), once we zoom out to the systems level to examine patterns of behavior among a broad range of actors, we might have introduced too many factors and too much uncertainty to allow for a traditional experimental design.



Examples of the dilemma at Moore

Health Care Delivery: In a field crowded with actors and funders working on quality of care, payment reform, and delivery system redesign, do Moore foundation's investments cause system changes that these same actors wouldn't have accomplished eventually anyway, perhaps by cobbling together funding from other sources? When many are working towards improving the quality of care and reducing medical errors, how can we untangle which specific results were enabled by the Foundation?

Science: Moore is rarely the sole funder for any of the science grantees. Almost all grantees receive funding from federal and private sources in addition to their Moore foundation funding. How might foundation staff parse out whether scientific advances made by grantees are due to Moore foundation funding versus other sources? How do we determine whether Moore Foundation funding accelerated the rate of progress even in the midst of other funding sources?

Systems change evaluation requires a shift in mindset about the goal of evaluation and the "fit" of different study designs to the situation at hand. If experimental designs to isolate causality are not possible, what alternatives are there? If we cannot, in most cases, attribute change to a single actor or cause, what standard of evidence should we be aiming to establish with our evaluation designs?

Potential Solutions

Our proposition at the Forum was that evaluations of systems change efforts require rigorously testing whether a plausible case for contribution can or cannot be made, rather than attempting to attribute a systems-level change to a single intervention or initiative. Some of the guest evaluators presented quasi-experimental and non-experimental study design options that aim to test the case for contribution by exploring and eliminating alternative hypotheses. These, and others that may be useful but were not presented at the Forum, include:

Process Tracing¹: A case-based approach to causal inference which focuses on the use of clues within a case (causal-process observations) to adjudicate between alternative possible explanations. Process tracing involves four types of causal tests, and can be used both to see if results are consistent with the program theory (theory of change) and to see if alternative explanations can be ruled out. *Presented at the Forum by [Jewlya Lynn](#) of Spark Policy Institute*

¹ All descriptions of methods, tools, and approaches in this paper are quoted or paraphrased from www.betterevaluation.org where possible, or other introductory sites, with the relevant page hyperlinked to the text so that staff can easily learn more about it. In addition, our favorite overview articles are included in the resource compilation. Finally, the evaluators who presented on a particular topic are linked to the description of that topic in case staff want to request more information about the method, tool, or approach.



and [Sarah Stachowiak](#) of ORS Impact, used to evaluate the contribution of the Collective Impact model across multiple sites.

Modus Operandi: A theory-based approach in which the evaluator builds a picture of predicted effects and supplement this picture through research. The combined picture helps the evaluator look for typical patterns of effects and contrasting patterns. The consistency of the 'trace' with the predicted pattern helps prove the program theory. Any differences from the predicted pattern helps disprove the program theory or open a new line of questioning. *Presented at the Forum by [Laura Leviton](#) of the Robert Wood Johnson Foundation, used to evaluate anti-smoking interventions.*

Ripple effect mapping: a participatory method that is best conducted for in-depth program interventions or collaborations that are expected to produce broad or deep changes in a group, organization, or community and to have "knock on" effects that are important to understand. Using elements of an appreciative inquiry approach, directly affected stakeholders and other critical observers brainstorm and hierarchically map the effects or ripples of the intervention, so that groups hypothesize about connections among observed program effects. Data is coded and themed, and connections between effects can be tested with additional data collection. *Presented at the Forum by [AnnJanette Rosga](#) of Informing Change*

Contribution Analysis: An approach for assessing causal questions and inferring causality that reduces uncertainty about the contribution the intervention is making to the observed results through an increased understanding of why the observed results have occurred and the roles played by the intervention and other internal and external factors. It entails assembling confirming evidence first and then testing its plausibility with other actors, and then assembling disconfirming evidence or evidence for alternative explanations for change.

Qualitative Comparative Analysis: A means of analyzing the causal contribution of different conditions (e.g. aspects of an intervention and the wider context) to an outcome of interest across a number of cases. QCA starts with the documentation of the different configurations of conditions associated with each case of an observed outcome. These are then subject to a minimization procedure that identifies the simplest set of conditions that can account for all the observed outcomes, as well as their absence. The results are typically expressed in ordinary language or as Boolean algebra. QCA results are able to distinguish various complex forms of causation (e.g., partial causes, necessary but not sufficient, etc.), and can be done with relatively small data sets.

Outcome harvesting: A multistep *inductive* evidence gathering process designed for situations where strategists have a high level of uncertainty about what outcomes might occur as a result of an innovative or highly emergent set of interventions. Unlike some evaluation approaches, Outcome Harvesting does not measure progress towards predetermined objectives or



outcomes, but rather, collects evidence of what has changed and, then, working backwards, determines whether and how an intervention contributed to these changes through a rather detailed and intensive evidence gathering process, which could include qualitative and/or quantitative data. The outcome(s) can be positive or negative, intended or unintended, direct or indirect, but the connection between the intervention and the outcomes should be plausible.

Propensity-score matching: When an intervention is more targeted to change the behavior or attitudes of a clearly-bounded audience (e.g., fellowships to cutting-edge researchers to encourage collaboration), but those targeted individuals cannot be randomly assigned to a control group and have some unique characteristics (such as strong leadership in the field), a quasi-experimental design might still be possible. In these cases, propensity-score matching can be used to estimate the difference in outcomes between beneficiaries and non-beneficiaries that is attributable to a particular intervention. A propensity score is an estimated probability that a 'unit' (in this case, a researcher) might be exposed to the program; it is constructed using the unit's observed characteristics. The propensity scores of all units in the sample, both beneficiaries and non-beneficiaries, are used to create a comparison group with which the program's impact can be measured. By comparing units that do not participate in a program, but otherwise share the same characteristics as those units which have participated, PSM reduces or eliminates biases in observational studies and estimates the causal effect of a program on an outcome or outcomes.

Presented at the Forum by [Carter Epstein](#) of Abt Associates, used to evaluate NSF's PIRE program to increase participation of US researchers in international collaborations.

Points of discussion and disagreement at the Forum

Moore staff experience the challenge of rigorously assessing contribution in a variety of ways, including:

- Understanding the contribution of Moore's weight and influence in a field (not just what activities its dollars directly "purchased"). For example, did Moore's entry into an area of science research result in a flooding in of other new funders? Did it raise the profile of the area of work and attract new young scholars even if they didn't receive funds directly? Some staff pointed out that this kind of contribution to change is not as highly valued at the foundation (and thus rarely captured in indicators and evaluation), but may in fact have a significant impact on either the likelihood that the larger-scale outcome they desire eventually occurs, or on the durability of that outcome over the long term.
- Understanding how to parse out whether Moore contributed the piece that was needed to get to critical mass or to catalyze the larger change. This is of course easier when you are the only funder, but much harder if you're working in partnership.

Evaluators suggested that establishing these kinds of contributions is easier when three things occur:

- Staff have a clearer theory up-front about the leverage points their strategy will target and the ways in which they will influence other actors in a field.



- Staff and evaluators do better at tracking an on-going timeline of key events (not just those initiated by the Foundation) that are influencing behaviors in the system. For example, did the Moore Foundation's announcement of a new pool of funds for investigator collaboration precede another funder's announcement? It can be difficult to reconstruct the sequence of critical events (a crucial aspect of testing contribution) after the initiative has been in place for several years.
- Evaluations (such as some of those suggested above) and should cast a wide net to find the full range of outcomes--predicted or not predicted-- that have occurred, then trace backwards to find both confirming and disconfirming evidence and to test alternate explanations for what triggered those outcomes. This is a different approach and mindset than the conventional one, where evaluators exclusively look for data on the indicators identified in advance as important to staff or the board.

One source of conflict that surfaced during the Forum was about the value of evidence produced by different kinds of study designs. Many research scientists and evaluators believe that only experimental designs can prove cause-and-effect and are thus the only approach worth investing in. However, a methodological model based on the assumption that counterfactuals are the only possibility for rigorously understanding causality is often unrealistic for most larger scale systems change efforts. An alternative view presented by some of the guest evaluators and Moore staff is that the purpose of evaluation is, in fact, to *reduce uncertainty* so that program implementers and funders can make better choices. In this context, both the standard of evidence is different and the range of study design options that may be useful is broader.

Suggestions for future learning and discussion

Establish principles of evidence. The Moore Foundation could continue a foundation-wide discussion about what kinds of evidence are both expected and possible, given the nature of the strategies and the level of investment in data collection and evaluation. Staff have different perceptions about what standards they are expected to meet and what kinds of data will pass muster with leadership and the board. Clarity and transparency on this issue could improve the screening and selection of evaluation proposals, increase the sense of fair treatment across program areas, and help staff judge whether their evaluation work is on the right track. Given resource limitations and the realities of data collection, consider setting a bar for evidence that is “good enough” to reduce uncertainty and lend increased confidence to decision making

Build capacity and knowledge on methods and design. Consider additional learning and capacity building on qualitative methods, mixed method designs, and non-experimental designs, so that staff have a wider range of options to choose from and can create a better fit between the nature of their strategies and their evaluation approach.

Test and learn from new approaches to establishing contribution. Upcoming evaluations are prime opportunities to experiment with evaluation designs that establish contribution in different ways. Consider asking applicants to propose rigorous methods for assessing contribution, such as process tracing, contribution analysis, QCA, outcome harvesting, etc. Create a small working group of program and MEL staff whose programs are undergoing these kinds of evaluations to compare notes on what it takes to get findings they have confidence in. For initiatives and portfolios facing mid-term evaluations, consider using some of the same principles demonstrated in these methods to reduce the risk of confirmation bias.

Start external data collection earlier. Consider contracting with evaluators or using in-house MEL staff, if you are not already, more robustly at the outset of an initiative planning process so that they can: a) help build thorough “baseline” snapshots of the system and its dynamics, b) help set up on-going tracking processes that will result in better data for a retrospective evaluation at the end of an initiative, or c) support on-going investigation of and learning about interim outcomes and changing dynamics throughout the course of an initiative.

Dilemma #2: How do we select and capture meaningful interim outcomes to understand systems change progress when we are operating in short time horizons and using strategies that have indirect impact?

The nature of the dilemma

As illustrated in the SEEP Kenya case discussed at the Forum, as well as in the WSEI case discussion, systems change efforts are characterized by “indirectness of impact.” For example, patient safety



initiatives ultimately care about health outcomes, but they seek to improve health outcomes indirectly by influencing how the systems and actors that deliver care operate. Market systems approaches to conservation ultimately aim to produce healthy and productive fisheries and forests, but they do so by influencing the behaviors and operations of the market actors that make use of those resources. In other words, system changes efforts are rarely directly impacting their ultimate “target audience.” In fact, the ultimate changes these efforts hope to produce often won’t materialize until long after significant changes in other actors occur. And yet, the change in the target audience is often the lone outcome to which these kinds of change efforts are held accountable.

When staff are working towards outcomes that will likely take years or decades to materialize, but are judged on their progress in 3 to 5-year cycles, what should they be measuring to understand progress? Additionally, when the larger context is shifting around them—economies boom and bust, who has political control changes with each election, major funders come and go from the system—sometimes progress will backslide and indicators will look worse before they look better. Or some aspects of the system will improve while others worsen. In many foundations, program staff are managing all of this shifting terrain while boards are looking for steady upward-trending progress on a small handful of narrow quantitative outcome indicators. Given this situation, how can staff select meaningful interim outcomes and indicators that give a reasonable sense of assurance that the system is headed in the right direction and that they are making a difference?

According to discussion at the Forum, this dilemma manifests in four ways at the foundation:

- 1) Fit of outcomes: What kinds of outcomes are appropriate for different stages of a systems change initiative that aims for “indirect impact”?
- 2) Package of outcomes: How can outcomes work together as a package to tell a more comprehensive and meaningful story to different audiences about what’s happening (particularly to staff, who need data for strategic guidance, and the board, who tend to be more comfortable with seeing steady increases in the indicators for a target outcome)?
- 3) Interpretation of outcomes: How can we increase our confidence that an interim, indirect outcome is actually *important* to achieving our long-term goal?
- 4) Evolution of outcomes: How can we identify and track outcomes in a way that evolves with our understanding of how the system operates and what needs to change, rather than following a fixed set of outcomes even when we discover they may not be relevant?

Examples of the dilemma at Moore

Though the strategy isn’t fully fleshed out yet, we can imagine a challenge in patient safety work, where initial outcomes may include a change institutional appetite, capacity for, and agreement on a set of



practice changes. Later outcomes may focus on successful implementation of a new practice in a few places as proof of concept, and increases in political will for shifting payment system incentives. The time it takes to achieve any of these interim outcomes at a system-level scale may change considerably depending on what happens to health reform. And changes to health reform may in fact cause indicators on the ultimate *raison d'être* for the initiative (improved health outcomes) to worsen before they get better.

In some cases, the core outcome of an initiative is, in fact, directly funded by the Foundation, but the long-term durability of the outcome depends on the initiative also being able to generate impact indirectly on other actors in the system. For example, the EPIQS initiative provides resources directly to the field's top scientists for basic discovery-driven research in materials synthesis, and to young experts to become established at US academic institutions. It also directly funds theory centers and instrumentation.

While these investments directly enable the scientific discoveries the Foundation cares about, it is possible that the outcome will be fleeting if the universities don't also develop an appreciation for materials synthesis and decide to fund it after the Moore Foundation no longer does. In this case, measuring the outcomes related to shifts in the academic system's commitment to the field (e.g., a long-term budget, permanent appointments or cross-disciplinary centers, a growing pipeline of young scholars, academic journals and annual conferences dedicated to the topic) would be critical to understanding the full value and impact of the EPIQS approach.

Potential Solutions

Suggestions offered at the Forum for managing this dilemma address these four components of the problem to different degrees.

✓ **Use and adapt existing systems change frameworks**

First, we proposed that staff explore existing systems change models and frameworks to 1) plan where the strategy should focus, 2) anticipate how outcomes and indicators might evolve over time, and 3) identify the range of outcomes that could be of interest at different stages. Guest evaluators offered some frameworks they have found useful:

[Disrupting Systems Dynamics framework](#) – A framework for market systems interventions that outlines a pathway for systemic change, and presents a complementary set of indicator areas, or 'domains', that signal systemic changes. The framework distinguishes between agent-level change (changes in voice, level of investment, amount of innovation, perceptions and beliefs, imitation, and institutionalization of new practices) and collective level change (changes in norms, governing rules, and networks)

Created and presented at the Forum by [Ben Fowler](#) of MarketShare Associates.



ABLE Change Framework – A model designed to help communities and states more effectively address social issues affecting children, youth, and families. It helps community collaboratives unpack the domains of a system that likely need to change (a lá indirect impact) to create a system that produces the ultimate desired outcome. These domains are: mindsets, power, resources, regulations, connections, components, and the interactions between all of these. Initiatives may focus directly on changing only a sub-set of these but see ripple effects in other domains. Alternately, they may sequence their focus to address different domains of the system over time.

Created and presented at the Forum by [Pennie Foster Fishman](#) of Michigan State University.

Consolidated Framework for Implementation Research – A framework offering a range of constructs that are critical to the successful implementation of a program or approach. These include characteristics of: the intervention itself (adaptability, cost, relative advantage, etc.), the inner organizational setting and outer contextual setting within which the intervention takes place, the individuals involved, and the implementation process itself (quality of planning, engaging, evaluating, etc.). Though the model was developed to improve the quality of health service delivery, many have found it a useful frame for thinking through larger scale system redesign processes and interim outcomes in other content areas.

Presented at the Forum by [Steven Asch](#) of Stanford University and the Center for Innovation to Implementation.

As staff explore and adapt existing frameworks, it is critical that they consider the unique mechanisms by which scale or systemic change (as opposed to single-actor or temporary change) happens within the system they are addressing. For example, in market systems, “crowding in” to a new market and imitating successful competitors are critical mechanisms by which a particular behavior spreads among market actors. Other systems, such as health care and K-12 education, more commonly see spreading or scaling through formal rules, policies, funding/payment requirements, and professional development and training standards. One table discussion at the Forum, led by [Clare Nolan](#) and [Julia Coffman](#), explored various theories for scaling impact in order to tighten and clarify thinking about how different mechanism for scale might be articulated and measured.

✓ **Consider using predictive methods to “pressure test” the value of interim outcomes**

There are a range of qualitative and quantitative methods that predict whether future changes might occur. Several of these can be modified to help increase our confidence that a particular set of interim outcomes improve the chances that a future outcome will be achieved.

Limiting Factor Analysis – a qualitative approach to forecasting developed to determine the likelihood that biodiversity conservation projects will succeed, given the degree to which they address the factors that present the greatest threat to the project. It entails conducting a system analysis to identify limiting factors, such as institutional capacity, public policy, enforcement practices, stakeholder support, etc., and inviting a range of experts and



stakeholders to score and rank the current status of the limiting factor. If done at multiple points throughout an intervention, the method can help to assess whether the greatest barriers are being addressed over time and/or whether new barriers have arisen, and can help point to new areas for intervention that could increase the likelihood of success.

Developed and presented at the Forum by [Jared Hardner](#) of Hardner & Gullison Associates

Scenario Planning – common in the private sector to evaluate strategic choices, scenario planning is also a qualitative predictive exercise that can be applied to evaluation. It involves first identifying a full list of immediate factors and remote forces that could affect the outcome of interest and then prioritizing the top 2-3 based on their degree of importance to the outcome and the amount of uncertainty about the factor. Teams flesh out possible futures where these forces play out in different ways. To apply to evaluation, teams judge the extent to which particular interim outcomes increase the systems' ability to be resilient and make progress in the face of different possible futures. Drawing on a wide range of experts and insiders from multiple perspectives in this process reduces the risk of confirmation bias and increases the plausibility of the conclusions. Interim outcomes that are meaningful to the ultimate outcome in all scenarios may be most critical, while those that are most vulnerable to different futures are more tenuous signals of progress.

Presented at the Forum by [Rhonda Evans](#) of the Monitor Institute

Agent-based modeling – A quantitative/computational model for simulating the actions and interactions of individual actors or groups to determine their potential effect on the system as a whole. Most ABM is done via software and requires a team to conceptualize numerous agents specified at various scales; name the decision-making heuristics that guide agent behavior, define learning rules or adaptive processes; and set other parameters. Once the model is built, teams can make small or large adjustments in parameters to see how those changes would affect other actors and the whole system. ABM has been used to predict the effects of certain interventions in health care, urban planning and infrastructure investment, economic policy, and many other fields, it can also be used to simulate the future impact of interim outcomes. By setting parameters or changing decision making rules to reflect actual measured changes (i.e. interim outcomes) in norms or investment priorities of other actors, for example, teams can forecast future impact on other outcomes. The upside of agent based modeling is that the software is capable of handling non-linearity and multiple feedback loops in ways the human brain is not. The downsides are that it can be very costly and time intensive to build a quality, sensible model; it's very easy to inadvertently leave out critical parameters that would dramatically change the performance of the model; and the results can be difficult to interpret.

✓ **Avoid creating rigid or overly detailed outcome frameworks at the outset.**

In our experience, during the initiative planning process many foundations and evaluators are in the habit of creating a very detailed outcome framework that stretches through the course of a multi-year initiative and then never changing it. While the discipline of thinking through the interim outcomes can produce stronger plans, over-doing it can be a waste of time. Smart systems



strategies will *necessarily* adapt over time in response to changing dynamics and new insights about how the system works. Fixed, detailed frameworks can become irrelevant quickly. Instead, consider beginning with a handful of outcomes and indicators that are closely linked to the initial stages of work and then build in a regular time to re-visit, flesh out, remove and add outcomes and indicators as the initiative evolves.

✓ **Avoid presenting and interpreting individual outcomes and indicators in isolation.**

While no session focused specifically on this issue, several participants at tables and large group discussions noted challenges with taking a “portfolio” approach to measurement and evaluation. In many cases, individual grants and interventions within a portfolio are intended to have a synergistic effect, producing together a change that is greater than the sum of the individual parts. Other portfolios are intended to cover a mix of change levers, with the understanding that in complex systems, some interventions will inevitably fail to gain traction or some outcomes will move backwards in the face of external forces. Aiming for a mix of outcomes that are additive, that account for risk and unpredictability, and that combine long shots with sure bets are good strategic practices for systems change work. But it goes against convention in much of philanthropy, where we have established habits and artifacts, such as prioritizing a single outcome as a guiding “north star” or creating dashboards that reduce the number of indicators we use to communicate progress to only a small few, often presented with little context.

One solution shared briefly at the Forum included: Use packages of outcomes and indicators, always created, presented, and interpreted together and in context. Imagine them like a bundle of vital signs that together describe the overall health of the “patient,” that are interrelated but might move in different directions. As with a person’s health, no one indicator is dispositive. This approach requires more time and energy spent among staff and with the board interpreting data and exploring its implications in context rather than “letting the data speak for themselves.” In other words, the data ought to be the starting point for discussion and meaning-making, rather than the end point.

✓ **Embed interim outcomes in a research-based model or theory of change to illustrate the connection to longer term outcomes.**

Often interim outcomes are provided without clearly illustrating that they are part of a larger theory. Several people offered research-based theories or frames that staff can use to illustrate that they are at a critical mid-point in a long pathway and to make the reasoned case for why it’s important and what remains to be done.

- John Mayne’s “causal package” framework considers the degree to which your indicator package includes a strong mix of ground preparing, triggering, and sustaining factors.
- “Punctuated equilibrium” theory of change helps to illustrate tipping point interventions.
- ORS Impact’s [10 Pathways to Policy Change](#) offers research-based theories of policy change.



Points of discussion and disagreement at the Forum

- Many staff would like more focus on capturing unintended outcomes, both positive and negative, but have the sense that the methods available to do this (e.g., ripple effect mapping, outcome harvesting, appreciative inquiry, etc.) feel soft and subjective. There is a point of tension around whether participatory methods such as these have real value and fit with the culture of the institutions the initiatives are working with.
- The selection of outcomes and indicators can be very political and play into existing power dynamics. Some actors in the system may not want data on particular outcomes to be available to other actors. Decisions about measures to monitor the health of a system as a whole should take into account who has power, who gets to define success, and what their motivations are. This raises concerns for staff about when the foundation should set its own measurement frameworks vs. when the frameworks have to be developed collaboratively with other actors in the system. How might initiatives get more buy-in from grantees, partners, and other system actors into what to measure? How well do the foundation's expectations and processes allow for a collaborative measurement approach?
- Not all program areas have the same habits of doing robust systems analysis during the planning phases. This limits their ability to identify the full range of outcomes they hope to see, or to articulate ways in which their portfolio as whole might affect the larger field. For example, several staff observed that they haven't done enough to flesh out the scaling/spreading mechanisms in their strategy, and thus have few outcomes and indicators that help them understand whether it is occurring. Others find the "skinny" version of planning and articulation appropriate to the nature of the work and prefer to think about their work on a grant-by-grant basis. This tension is likely to affect the take-up of the initiative lifecycle work.
- Staff are concerned with the concrete logistical and resource challenges of collecting data on a sufficient range of systems level indicators. How interested is Moore in committing the resources necessary to support that kind of data collection? Is the trade-off worth it? How might initiatives do so without putting too much burden on individual grantees?



Suggestions for future learning and discussion

Consider cross-program learning on systems thinking tools for planning, such as identifying leverage points, understanding system dynamics, etc. Those who have already done this work can serve as case examples, illustrating advantages, pitfalls, and the concrete ways in which systems thinking affected their strategy. To avoid staying too conceptual, invite specific initiatives to “workshop” their grantmaking through these tools to draw conclusions about where they are/are not applicable and to right-size planning process expectations to different content areas and strategies. Without more robust systems analysis on the front end of planning, many of the problems with outcome and indicator selection cannot be solved.

Test out “light touch” future-forecasting methods on initiatives that are in the mid-point review phase. Consider bringing midpoint evaluation results and/or indicator data into a predictive exercise with a range of experts to rate their importance or relevance to the longer-term outcome. Present these indicators to the board with a summary of judgements about which are critical/sufficient, which are less so, and which indicators need to improve significantly in order achieve the longer-term goals.

Explore the board’s appetite for building their own knowledge about indirect impact and outcome frameworks for system change. Because boards often drive an organization’s perspective and approach to measurement, it may be beneficial to provide the board with some outside resources on systems-change outcomes. Learning how to interpret packages of indirect outcomes requires a different mindset and skillset than most are accustomed to.

Compile a bank of systems change frameworks that staff can use and adapt as they plan or design monitoring and evaluation for their initiatives. Share systems and outcomes frameworks between initiatives and programs to generate ideas and invite pressure testing.



Dilemma #3: How do we set realistic boundaries around the aspects of the system that our strategy and our evaluation should focus on?

The nature of the dilemma

Complex systems often have “fuzzy” boundaries. This means that they interface and interact with their larger environment by receiving inputs and delivering outputs external to their boundary. These boundaries permit the exchange of materials, energy, information or ideas. Practically speaking, this means that the range of factors that affect how a system operates (and thus the outcomes it produces) could be enormous. For example, the functioning of the health care system is affected by characteristics and actors directly participating in the system (payers, hospitals, medical schools, patients) as well as clear system rules and norms. But its operation and outcomes are also affected by exogenous factors, like health of the economy and employment rates, income inequality and social determinants of health, and even seemingly unrelated issues like immigration policy.

Boundary setting requires determining what factors are endogenous and exogenous, what things will be excluded from the analysis of the system, what factors are too far afield to try to address with your strategy, and what factors the evaluation should pay attention to. Boundary setting also requires setting a time horizon; how far back into history should we look to understand determining factors? How far forward should we look to predict patterns and outcomes? People who are newer to systems change work often find themselves creating vast, unwieldy systems models or maps in an effort to fully understand the area they’re working in. This can be paralyzing, as it offers too many options for intervention points and too many things to track. And for program strategists, as they “pull the thread” of one particular factor, they can easily find themselves working on distal factors that feel further and further away from direct impact. But how can they know when they are addressing a fundamental root cause that may seem too distant, but is in fact a significant leverage point, versus when they have gone too far afield?

Examples of the dilemma at Moore

This dilemma could easily appear in the health program area. But even in more clearly bounded systems, such as academic institutions, the decision about where to draw the line can be difficult. For example, should initiatives aiming to stimulate cross-disciplinary collaboration in basic science discovery be concerned about the role of the tenure system in rewarding individual scholarship and accolades? Should they try to address falling state budgets for higher education because cuts will likely affect cross-disciplinary institutes and centers and force scientists doing discovery work to vie for external grants that pull them away from a discovery approach?

Potential Solutions

In general, boundary setting is a matter of experimentation and judgement, as well as the simple reality of what the board will accept. Some systems thinkers advise listing out every influencing factor and



then mapping their relationships to one another using a mapping approach such as causal loop diagramming that shows the interrelationship between drivers. However, this approach can often lead to elaborate ‘spaghetti’ diagrams that no one can interpret or use. Others suggest using a winnowing process to focus on those elements that have direct line of sight to the ultimate goal and that are within the foundations’ scope of influence. Either way, most suggest that boundaries should be periodically revisited to assess whether, through the course of the work, you have learned that some critical “close in” factors can’t be unstuck without addressing more distant factors.

While the Forum did not offer many concrete tools specifically for boundary setting in either strategy design or evaluation, there are some approaches that can help.

Critical Systems Heuristics – a framework of 12 “boundary” questions designed to outline and provoke systematic thinking about boundary judgements. These questions, which probe sources of motivation/incentives, power, legitimation, and knowledge in a system are answered in terms of what currently is and what ought to be. This helps strategists determine who should be in their frame of reference as they examine how the system works and where to intervene.

Attractor Mapping – operating much like a heat map, attractor mapping helps to distinguish areas of intense activity (and thus opportunity or energy for change). Activity or energy can be overlaid onto maps of relationships (such as social network maps), organizational roles (organization charts), system dynamic models and causal loop diagrams. While attractor mapping does not help determine system boundaries, it may serve as another layer of information to support decision making about where to intervene that could capitalize on current momentum.

Created and presented by [Cameron Norman](#) of Cense Research + Design

Finally, guest evaluators suggested a systematically identifying which exogenous factors should be tracked throughout the initiative because the direction they move will affect the degree to which you can move other levers in the system. A scenario planning or pre-mortem approach can help identify the exogenous factors to which the initiative or the system as a whole is most susceptible. In other words, setting too tight a boundary on your *evaluation* by focusing exclusively on metrics that are directly tied to your specific interventions may cause you to misunderstand why something isn’t working or what’s driving progress.

Suggestions for future learning and discussion

Test the Critical Systems Heuristics framework on an initiative that’s struggling with boundary issues. Share results across teams so they can see how it was used and how it affected decisions.

Include more explicit descriptions of what falls outside the boundary in your strategy development process. Oftentimes, boundaries are only implicit. The practice of naming it more explicitly is a good way to reveal underlying assumptions about drivers and opportunities.



Dilemma #4: How can we embed continuous, intentional learning into our strategy and evaluation work?

The nature of the dilemma

Retrospective evaluations to support decision making at the board level offer staff little information for real-time strategic adaptations. As described in the Kenya SEEP case, complex systems change requires tools for smart sensing and navigation, as well as a habit among program staff of regularly naming assumptions, testing them in action, and adapting according to what is learned. However, staff managing multi-grantee initiatives with lots of moving parts often struggle to find the time and space to systematically reflect. They also often find that data that is collected to satisfy grant monitoring needs provides little real insight into what's happening. Finally, organizations that place a strong emphasis on accountability in the form of hitting targets often inadvertently create cultures that are inhospitable to admitting when things aren't working as hoped.

Building an organization's learning muscle often requires improvements in systems and processes that support the capturing, processing and interpretation of data that is actually useful for decision making. Much harder to address are informal attitudes, norms, habits of working, and the know-how to engage in genuine inquiry. We all learn constantly, but learning in teams and learning systematically in a way that accelerates progress is much more of a challenge.

Examples of the dilemma at Moore

Although many staff described a desire to learn more intentionally and systematically, few were comfortable offering concrete examples of ways in which learning is stymied at Moore. Instead we heard many general descriptions of barriers, which we have seen in many other foundations as well, including:

- Emphasis on retrospective evaluation to support the board but not staff, so that evaluations do not return data that is timely or relevant enough to on-the-ground choices that staff or grantees face.
- Few norms around on-going engagement with evaluators to support real-time data collection and reflection.
- Not enough time in the day to reflect (the most common answer at Moore and in a [benchmarking survey](#) of learning and evaluation practices among foundations).
- Strong internal competition and other organizational incentives to avoid admitting disappointments.
- Little opportunity to interact with colleagues and learn about their approaches.
- Knowledge/grants management systems that aren't quite robust enough to manage and mine for learning.
- Need to improve skills and systems for collating data, pattern recognition, and interpretation across such a broad range of interventions.
- Trouble gathering meaningful data on some of the tougher aspects of the work where we need to learn the most, such as changes in reputation, influence, norms, political will, collaboration, etc.



Perhaps most challenging is the large “learning culture” question: How can foundations integrate rigorous and robust shared learning into their ongoing habits and ways of working, such that it is an indispensable part of the work of grant making?

Potential Solutions

A considerable amount of thinking and work has been done on learning in philanthropy and learning in systems change efforts in particular. While learning individually can be an intuitive process, learning in teams requires more intentional structure and process to make thinking and assumptions visible, draw meaningful insights that can be shared and applied in different contexts, and store and pass along knowledge as staff change over time.

- ✓ **Build organizational incentives for learning, and diagnose and address disincentives.** Both formal and informal incentives often work against the development of more systemic learning habits. Formal incentives could include building learning expectations in program staff job descriptions and performance reviews, requiring the inclusion of specific learning questions and plans in grantmaking strategy proposals, or offering larger budgets for those who want to invest in learning as a team or in methods for sharing learning outside the foundation. Informal incentives could include verbal recognition and reinforcement from leadership and trustees for staff who are more candid about challenges and who have robust learning strategies. Modeling curiosity, candor, and admission of mistakes by leaders can also go a long way towards shifting culture. Finally, regularly check the ways that processes and language (e.g., how the grant review process is structured and how the foundation talks about accountability) dampen people’s willingness to reflect deeply and openly. And because time is the most common barrier busy program staff cite, take a realistic look at workloads and time analyses to determine whether it is, in fact, possible and fair to ask staff to pause more frequently for systematic reflection without revising existing workloads.
- ✓ **Re-cast existing meetings, processes, and structures to foreground more intentional learning rather than treating learning as an add-on task.** Organizations that have made strides on learning often begin by revising basic meeting habits. Team or staff meetings that include updates or task updates are retooled around strategic questions. Meetings can be more thoughtfully structured to include outside/alternative perspectives to avoid groupthink, introduce different disciplinary lenses. There are myriad resources and tools available to help teams change the quality and depth of their conversation. They are all designed to help teams close the learning loop by tightly connecting insights drawn from the data to concrete implications for the work:

[Facilitating Intentional Group Learning: A practical Guide to 21 Learning Activities](#). Offered by FSG, this guide offers several ideas for designing and facilitating learning activities at varying levels of time and intensity for teams or groups. The guide includes a set of methods specifically useful for systems-thinking learning.

Developed and discussed at the forum by [Hallie Preskill](#) of FSG.



[After Action Reviews](#). A structured debrief process, already in use by teams at the Foundation, which helps teams reflect on their initial hypotheses about how to accomplish their goal, how and why the results played out as they did, and what the implications are for the next iteration. Teams that build a habit of conducting AARs regularly can accumulate and carry forward knowledge more effectively over time, becoming less likely to repeat mistakes. AARs are even more powerful when paired with [Before Action Reviews](#), which help teams clarify what they hope to achieve and anticipate what issues might arise during their work based on past experience.

[Emergent Learning Tables](#). This is a structured approach to learning over time in teams or groups about substantive strategy questions. It can help team develop a “learning muscle” by getting them in the habit of reviewing data and experience through the frame of a meaningful strategic question, drawing insights, generating forward-facing hypotheses about how to improve the work, and identifying concrete opportunities to test the hypotheses in action. Groups then observe what happens (hopefully with evaluative data) to determine whether their hypotheses bear out and begin the cycle again. Over time, teams that use the EL approach often get faster at identifying meaningful strategy questions that can be answered with data and better at transforming what they learn to concrete adaptations in strategy.

Developed by [Fourth Quadrant Partners](#).

[How Shortcuts Cut Us Short: Cognitive Traps in Philanthropic Decision making](#). Even when the right data are available and teams set aside time to reflect, behavioral economics tells us that our cognitive biases can get in the way of quality deliberation and decisions. This brief describes a set of cognitive traps particularly common to philanthropy and recommends very concrete approaches for guarding against them by adjusting the kinds of data brought to the table and the decision-making processes people use.

Developed and discussed at the Forum by [Julia Coffman and Tanya Beer](#) at the Center for Evaluation Innovation.

- ✓ **Treat learning as an indispensable part of the strategy—and as a critical outcome—from the outset of the strategy design process.** An increasing number of foundations are explicitly asking program staff to “bake” a learning stream into their strategy during the initial design phase. Program staff articulate how the team—and often grantees—will use data and evaluation to learn over time and to adapt other components of the strategy and/or to influence other actors. Learning as a strategy stream has its own budget line, key milestones, and sometimes even a clear articulation of the “theory of change” explaining why the learning approach is designed as it is and what outcomes it is intended to achieve. Just as a periodic strategy review would examine whether the strategy was proceeding as hoped and having the intended effects, it also examines whether the team and grantees are using data and evaluation for smart adaptation as intended. This approach is more robust when an evaluator is engaged at the outset of the strategy design process to help teams identify which questions might be most value-added to answer with data and to begin data



collection immediately. And when learning is viewed as a critical outcome of the work for which staff are accountable, they are more likely to commit serious time and energy to it.

- ✓ **Consider [developmental evaluation](#) where appropriate.** DE is an evaluation approach explicitly intended to support innovation, where people are working under conditions of high complexity and the way forward is uncertain. The evaluator's role is as a "critical friend" to the team, identifying and answering real time questions with rigorous data about what's unfolding so that the team can make decisions about next steps. Note that DE is not right for more stable initiatives where the way forward is fairly well known and agreed upon and adaptations are expected to be minor.

- ✓ **Hone the skill of asking the right questions.** Systems thinking expert and organizational learning scholar Peter Senge was once asked in a conference: what's the single most important skill for a learning organization?" His response: "They know how to ask the right questions." Organizations whose learning returns the most value are those who pose meaningful questions that, when answered, make a real difference in their ability to achieve the results they want. In our experience, people often organize and frame their learning retrospectively, e.g., "What have we learned about X?" without real consideration as to whether the answer to that question would affect how they work going forward. This makes it tough to translate retrospective insights into action in the future. Strategy questions are forward-looking and action oriented. "What would it take to...?" "How can we...?" Data and evaluation about work that happened in the past (even yesterday) is meaningful to strategy when interpreted in a way that helps to identify patterns and reveal insights about what's likely to work (or not work) in the future under particular conditions. It can also be fruitful to "game out" the kinds of data you might get back from a data collection effort or in response to a learning question. If the data said X, would you do something different? Would you know *what* to do differently? How about if it said Y? If not, you probably haven't landed on a meaningful learning question.

Suggestions for future learning and discussion

Determine to what end the foundation wants to improve learning and whose learning to prioritize.

Consider what signals might indicate that the learning efforts you're experimenting with are working and worth the time and resources. Systems initiatives may need to learn in a different way than field-building efforts or standard programmatic grants.

Start small. Rather than creating large-scale learning systems and foundation-wide processes all at once, try setting up a small voluntary working group of staff who will experiment with different learning methods for a period of time to see if it increases the depth of reflection and provides them with more strategy direction. Ask this group to learn together through experimentation what it takes to shift team habits around learning.

Sketch out the formal and informal incentives, norms, and mental models at Moore that either enable or impede learning practices. Try addressing a few high leverage ones, such as how program and foundation leaders model learning themselves and respond to staff who are transparent about disappointing results.

Focus on getting 2-3 concrete "wins" on high-return learning. People will carve out the time and resources for more intentional learning if they believe it will actually make a difference in their results. That means getting clearer about the difference between learning for learning's sake (after all, who isn't already always learning?) and learning in ways that increase impact. Having some high-return examples could help make the case to the board and staff for expanding its approach to learning.

Engage evaluators early in the strategy design process for the next systems change strategy on the docket. Having an ongoing relationship with an evaluator or MEL team member through the course of complex initiative is likely to help the staff with its strategy questions AND result in higher quality data and findings for the board.

Consider some professional development or a community of practice for MEL staff or program staff who are interested in supporting more learning in their teams or with grantees. Organizational learning is a skill and a capacity that has to be developed and practiced over time, and it is distinct from evaluation expertise.

EXTERNAL OBSERVATIONS ABOUT ORGANIZATIONAL CULTURE AND LEARNING

Many of the evaluation dilemmas raised by Moore staff are common to any foundation that invests in large-scale systems change initiatives. However, there are unique aspects of the Moore Foundation culture that affect how these dilemmas play out. They are worth mentioning because they will affect which measurement and evaluation solutions, approaches, and frameworks are likely to "stick." *Please note* that these are generalized outsider observations from a handful of meetings and interviews with program and MEL staff and are certainly over-simplified. Nor are they intended as criticisms—every foundation has a unique organizational DNA that affects what kinds of learning and evaluation approaches will be a good fit. They also do not capture all the nuances of the organization but are



simply intended to raise questions the foundation should consider as it continues to develop its MEL approach.

➤ **Commitment to Measurable Results**

The foundation has a deep, explicit commitment to measurable results. Its push to measure and quantify is driven not from a desire to take credit, but rather from a desire to have a meaningful impact—to use resources where staff have a level of certainty that they will make a difference. As one program officer said during the Forum, “We don’t fund causes, we fund results.” Unfortunately, the philanthropic sector offers too many examples of evaluation and measurement practices that rationalize nearly any outcome as worthwhile progress. This has, perhaps rightfully, resulted in suspicion among some at Moore that the language of complex systems (such as uncertainty, unpredictability, etc.) simply gives people an easy way out of being accountable for results. However, evaluation and measurement that *over*-presumes stability and predictability can create perverse incentives, such as selecting interventions that may be less powerful because their results will be more easily predictable and quantifiable. It can also create a false sense of certainty, inadvertently allowing another easy way out of real accountability (e.g., this indicator is going up, so we have been accountable to our results, irrespective of whether that indicator will go down again when your funding goes away or whether someone else was more responsible for the change). In order to maintain their commitment to results, the program staff, MEL staff, and external evaluators working with Moore should design measurement, evaluation, and learning approaches that actively battle against confirmation bias, while still acknowledging that the realities of systems change affect what kinds of study designs are appropriate and the degree of certainty that is possible. As MEL practices develop and evolve, it may be helpful to consider other mechanisms—such as rigorous on-going learning and assumption testing—can also support accountability.

➤ **Evidentiary Concerns**

A commitment to measurable results often presumes that evidence regarding program successes and challenges needs to be quantitative; indeed, strong causal claims in research typically depend on statistical findings. However, many forms of evaluation, even in the hard sciences, rely on qualitative methods and data. Over-prioritization of quantitative evidence produced by a narrow set of research designs and methods, moreover, can lead to missing critical human or institutional factors affecting outcomes—particular as outcomes relate to systemic change (e.g., the politics of placing a telescope on a particular hill in Hawaii). In short, not all types of change lend themselves to examination by quantitative methods. Therefore, without incorporating and taking seriously qualitative methods, the Foundation may miss out on the deeper understanding of the range of changes happening *and why they are happening* that can come from a mixed methods approach. As a result, there is likely a need for some capacity building on how to determine when qualitative methods and non-experimental designs are appropriate, how to judge their rigor and quality, and how to interpret them in the context of discussions about impact.



➤ **Academic/Business Unit Culture**

In many ways, the culture of the foundation mirrors academic or “independent business unit” culture, where faculty/units operate very independently, are careful to stay out of each other’s “lanes,” and prioritize high levels of content specialization (as compared to, say, foundations where staff come from the nonprofit management sector and thus tend to value a deep understanding of nonprofit capacity and performance management). It is an extraordinary asset that staff bring deep knowledge of their content areas and are likely to engage in intense learning with others who are within their specific field. However, this kind of deep content specialization can also cause staff to avoid asking one another tough questions or introducing alternative perspectives for fear of stepping on one another’s toes or sounding uninformed. The implication here is that the MEL staff cannot assume (as we discovered in this forum) that program staff will “dig deep” with each other or be comfortable challenging one another’s thinking, particularly in large groups and across program areas. A habit of robust peer-to-peer inquiry requires a level of trust and mutual curiosity that has to be developed over time. It also may require surfacing where different program areas face common or analogous challenges and doing targeted cross-program learning on those narrower challenges (e.g., some of the science initiatives are trying to change norms and incentives for collaboration across disciplines, and the health initiatives have long experience in incentivizing cross-institutional collaboration—a topic ripe for cross-programmatic learning).

Like almost all aspects of organizational culture, these characteristics are simultaneously assets for and barriers to the foundation’s ability to use MEL to increase its effectiveness. If MEL staff design measurement, evaluation and learning approaches that take these characteristics into account, there will be a higher chance to see real uptake. But they are also limiting factors that can block staff’s interest in using a broader set of approaches that are better matched to systems change efforts and real-time learning.

As a Foundation develops its approach to MEL, we recommend first building the capacity of program staff and board on the aspects of the MEL work that connect to the organizations most deeply held values and concerns. Otherwise, new MEL practices and systems can feel too threatening to the organization’s identity or too big a transition to manage. For example, program staff at foundations whose organizational DNA emphasizes equity and collaboration might feel that a foundation-designed MEL approach that focuses on non-systemic quantitative indicators reinforces the very power dynamics they’re working against. Instead, this foundation might get greater traction on MEL among their board and program staff if they start with by building a practice of shared reflection on personal experiences, collective sense making, and co-generation of evaluation questions with grantees. Over time, they can cultivate an appetite for more systematic and rigorous data to inform that reflection. Foundations whose culture is defined by a commitment to entrepreneurship and innovation can get traction in MEL by setting up light-touch flexible outcome frameworks using “provisional” language with a clear timeline for revisiting them, so that innovators don’t feel trapped in an inflexible course of action.



In the case of the Moore Foundation, early MEL efforts could tap into the Foundation's commitment to measurable results and standards of evidence by focusing on ways to: 1) guard against bias in planning, data collection, and interpretation of findings and 2) increase confidence in the relevance of interim outcomes. Consider partnering with program staff to review current data collection plans and strategy processes for opportunities to bring in alternative perspectives, to triangulate qualitative and quantitative data, and to articulate and compare sets of assumptions about how change might unfold. Try pairing a mid-term review process with forward-looking forecasting exercises so that staff experience a clear analytical connection between evaluation and planning. These approaches would tap into staff (and board) desire to deal with complexity without compromising their sense of accountability. If you expend too much time and energy developing large and elaborate MEL systems or frameworks without first demonstrating how to increase relevance and quality, the system may not "stick" or be perceived as adding any real value.

With respect to learning, consider focusing primarily on within-program-area learning (where people will be more likely to offer and accept alternative views and see the most immediate value for their work), but ask teams to slowly experiment with ways to bring in new perspectives both during planning and when making sense out of monitoring and evaluation data. Experiment with strategy review processes that invite a peer from another program team to play devil's advocate or pose tough questions to the team about their plans. Consider inviting in a handful of academics or practitioners from different disciplines to look at a problem, an approach, or a set of proposed outcomes from different angles. For cross-program learning, find opportunities among small groups to explore a shared strategic challenge through a range of disciplinary perspectives (e.g., Under what conditions are informal social networks a tool for change and how can influencers shape practice within their network? How are group/institutional norms formed, challenged or reinforced and what does that mean for our change strategies? How can we determine whether significant shifts in practices, norms or mindsets have reached a tipping point and/or are durable?), always asking about implications for strategy.

Finally, consider exploring with foundation leadership and program directors how new MEL mindsets and practices might be incentivized and reinforced at all levels of the foundation, including in the interaction between staff and board. How is candid learning and reflection—and the strategy adaptation that should accompany it—formally and informally rewarded? How is it modeled? To what extent do existing practices and processes send mixed messages about the purpose and philosophy of measurement, evaluation and learning at the foundation? And perhaps most importantly, where might small changes in MEL practices have the biggest impact on the foundation's ability to achieve its goals?