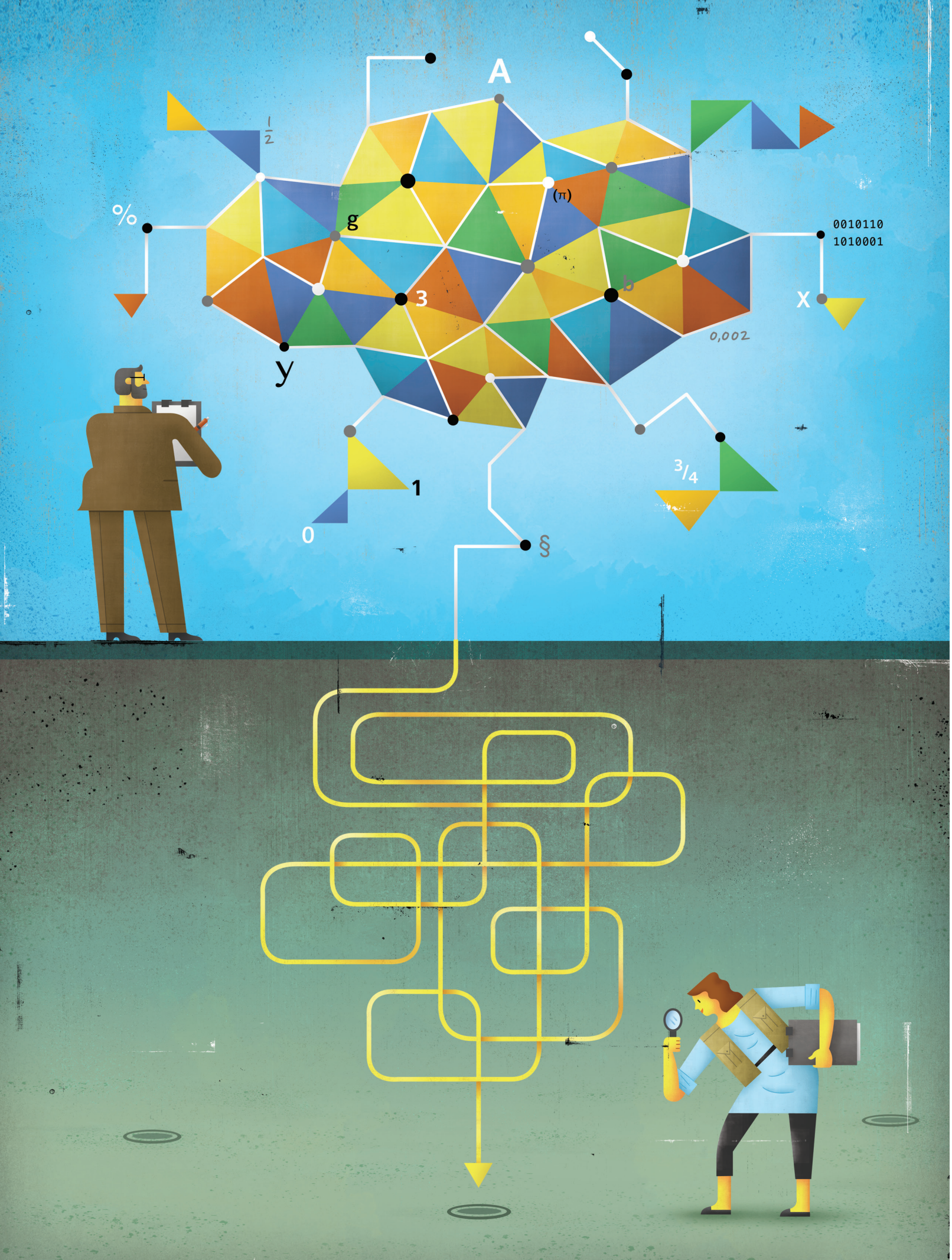
Stanford SOCIAL INNOVATION Review

Features

Ten Reasons Not to Measure Impact – and What to Do Instead

By Mary Kay Gugerty & Dean Karlan

Stanford Social Innovation Review
Summer 2018


A
1/2
%
g
(π)
0010110
1010001
3
b
x
y
0,002
1
3/4
0
§

Impact evaluations are an important tool for learning about effective solutions to social problems, but they are a good investment only in the right circumstances. In the meantime, organizations must build an internal culture in which the right data are regularly collected, analyzed, and applied to manage implementation and improve programs.

Ten Reasons Not to Measure Impact—and What to Do Instead

BY MARY KAY GUGERTY & DEAN KARLAN

Illustration by Davor Pavelic

Would you rather help one child a little bit today, or wait a few years and help five children even more? Every dollar spent on current programs is a dollar used to help today's children in need—a worthy cause. Yet every dollar spent on research today, in theory, is a dollar invested in helping tomorrow's children even more. Admittedly, this trade-off is complex, imprecise, and uncertain. But the promise of research that can help us do more good per dollar spent is enticing.

Yet here's one cautionary claim we can make for certain: Every dollar spent on poorly conceived research that does not help tomorrow's children is a dollar wasted.

Good impact evaluations—those that answer policy-relevant questions with rigor—have improved development knowledge, policy, and practice. For example, the NGO Living Goods conducted a rigorous evaluation to measure the impact of its community health model based on door-to-door sales and promotions. The evidence of impact was strong: Their model generated a 27 percent reduction in child mortality. This evidence subsequently persuaded policy makers, replication partners, and major funders to support the rapid expansion of Living Goods' reach to five million people. Meanwhile, rigorous evidence continues to further validate the model and help to make it work even better.

Of course, not all rigorous research offers such quick and rosy results. Consider the many studies required to discover a successful drug and the lengthy process of seeking regulatory approval and adoption by the health-care system. The same holds true for fighting poverty: Innovations for Poverty Action (IPA), a research and policy nonprofit that promotes impact evaluations for finding solutions to global poverty, has conducted more than 650 randomized controlled trials (RCTs) since its inception in 2002. These studies have sometimes provided evidence about how best to use scarce resources (e.g., give away bed nets for free to fight malaria), as well as how to avoid wasting them (e.g., don't expand traditional microcredit). But the vast majority of studies did not paint a clear picture that led to immediate policy changes. Developing an evidence base is more like building a mosaic: Each individual piece does not make the picture, but bit by bit a picture becomes clearer and clearer.

How do these investments in evidence pay off? IPA estimated the benefits of its research by looking at its return on investment—the ratio of the benefit from the scale-up of the demonstrated large-scale successes divided by the total costs since IPA's founding. The ratio was 74x—a huge result. But this is far from a precise measure of impact, since IPA cannot establish what would have happened had IPA never existed. (Yes, IPA recognizes the irony of advocating for RCTs while being unable to subject its own operations to that standard. Yet IPA's approach is intellectually consistent: Many questions and circumstances do not call for RCTs.)

Even so, a simple thought exercise helps to demonstrate the potential payoff. IPA never works alone—all evaluations and policy engagements are conducted in partnership with academics and implementing organizations, and increasingly with governments. Moving from an idea to the research phase to policy takes multiple steps and actors, often over many years. But even if IPA deserves only 10 percent of the credit for the policy changes behind the benefits calculated above, the ratio of benefits to costs is still 7.4x. That is a solid return on investment.

Despite the demonstrated value of high-quality impact evaluations, a great deal of money and time has been wasted on poorly designed, poorly implemented, and poorly conceived impact evaluations. Perhaps some studies had too small of a sample or paid insufficient attention to establishing causality and quality data, and hence any results should be ignored; others perhaps failed to engage stakeholders appropriately, and as a consequence useful results were never put to use.

The push for more and more impact measurement can not only lead to poor studies and wasted money, but also distract and take resources from collecting data that can actually help improve the performance of an effort. To address these difficulties, we wrote a book, *The Goldilocks Challenge*, to help guide organizations in designing "right-fit" evidence strategies. The struggle to find the right fit in evidence resembles the predicament that Goldilocks faces in the classic children's fable. Goldilocks, lost in the forest, finds an empty house with a large number of options: chairs, bowls of porridge, and beds of all sizes. She tries each but finds that most do not suit her: The porridge is too hot or too cold, the bed too hard or too soft—she struggles to find options that are "just right." Like Goldilocks, the social sector has to navigate many choices and challenges to build monitoring and evaluation systems that fit their needs. Some will push for more and more data; others will not push for enough.

To create a right-fit evidence system, we need to consider not only when to measure impact, but when *not* to measure impact. Given all the benefits of impact measurement, it may seem irresponsible not to try to measure it. But there are situations in which an insistent focus on measuring impact can be counterproductive to collecting other important data.

MISPLACED PRIORITIES

How have we reached this point? If impact evaluation is so important, why are we advocating for limiting its use? The rapidly decreasing costs of data collection and analysis have certainly helped to heighten the appeal of impact measurement. Thirty years ago, frugal budgets restricted long-distance calls. Now free videoconferencing can connect people from multiple countries all at once. Previously, organizations might have argued that collecting data is too time-consuming and expensive. Today, the cost of collecting, storing, and analyzing data is much cheaper. We can process millions of data points and spit out analyses to field operators in mere minutes. And the pace of change remains rapid: Satellite imagery and a multitude of GPS monitoring devices, for example, are rapidly influencing the way programs are run and the richness of the questions that evaluators and researchers can ask. Naturally, quicker and cheaper data also makes organizations and stakeholders more willing to demand it.

At the same time, there have been more calls for accountability in the public and social sectors based on this ability to more easily measure results. Major donor organizations from the Bill & Melinda Gates Foundation to the UK's Department for International Development (DFID) are requiring evidence of impact. Social impact bonds and pay-for-success programs seek to fund effective initiatives by tying financing to proven results. And proponents of effective altruism seek to persuade philanthropists to give only to programs with strong evidence of effectiveness.

The trend toward impact measurement is mostly positive, but the push to demonstrate impact has also wasted resources, compromised monitoring efforts in favor of impact evaluation, and contributed to a rise in poor and even misleading methods of demonstrating impact. For instance, many organizations collect more data than they actually have the resources to analyze, resulting in wasted time and effort that could have been spent more productively elsewhere. Other organizations collect the wrong data, tracking changes in outcomes over time but not in a way that allows them to know whether the organization *caused* the changes or they just happened to occur alongside the program.

Bad impact evaluations can also provide misleading or just plain wrong results, leading to poor future decisions. Effective programs may be overlooked and ineffective programs wrongly funded. In addition to such social costs, poor impact evaluations have important opportunity costs as well. Resources spent on a bad impact evaluation could have been devoted instead to implementation or to needed subsidies or programs.

Much of such waste in pursuit of impact comes from the overuse of the word *impact*. Impact is more than a buzzword. Impact implies causality; it tells us how a program or organization has changed the world around it. Implicitly this means that one must estimate what would have occurred in the absence of the program—what evaluators call "the counterfactual." The term sounds technocratic, but it matters a great deal in assessing how best to spend limited resources to help individuals and communities.

When feasible, the most straightforward way to create a counterfactual is through a randomized controlled trial (RCT) in which participation in a program, or in some aspect of a program, is decided partly through random allocation. Without a counterfactual, we do not know whether the program caused a change to happen or whether some outside factor—such as weather, economic growth, or other government policy—triggered the change. We can't know whether those who participated in a program changed their lives because of the program or because of other factors. A rigorous counterfactual can change conventional but misplaced beliefs: For example, recent counterfactual-based impact evaluations of microcredit programs found much lower impact on household income than was previously claimed by microcredit advocates.

Good monitoring data are often collateral damage in the pursuit of measuring impact. Information on what the staff is doing, take-up and usage of program services, and what constituents think of operations

MARY KAY GUGERTY is the Nancy Bell Evans Professor of Nonprofit Management at the Daniel J. Evans School of Public Affairs at the University of Washington and the faculty director of the Nancy Bell Evans Center on Nonprofits & Philanthropy.

DEAN KARLAN is professor of economics and finance at the Kellogg School of Management at Northwestern University, where he is codirector of the Global Poverty Research Lab at the Buffett Institute for Global Studies. He is also founder of Innovations for Poverty Action and cofounder of ImpactMatters.

Gugerty and Karlan are coauthors of *The Goldilocks Challenge: Right-Fit Evidence for the Social Sector* (Oxford University Press).

can help create a better program and stronger organization. These data often get lost or overshadowed in the pursuit of impact evaluations. This is partly understandable: impact is the ultimate goal, and sloppy thinking often conflates management data with impact data. (Take-up of a product like microcredit, for example, is an important piece of management data but is not a measure of impact; statements such as "50,000 clients served" do not measure impact.)

The challenge for organizations is to build and use data collection strategies and systems that accurately report impact when possible, demonstrate accountability, and provide decision makers with timely and actionable operational data. The challenge for funders and other nonprofit stakeholders is to ask organizations to be accountable for developing these right-fit evidence systems and to demand impact evaluation only when the time is right.

In what follows, we offer 10 reasons for not measuring impact. We then provide a framework for right-fit monitoring and evaluation systems that help organizations stay consistently and appropriately attuned to the data needed for accountability, learning, and improvement.

THE 10 REASONS

The 10 reasons not to measure impact fall into four categories: *Not the Right Tool, Not Now, Not Feasible*, and *Not Worth It*. For each reason, we also offer alternatives that fans of impact evaluation can adopt instead.

1. Not the Right Tool: Excellent question, wrong approach.

Here are some excellent questions you may ask in evaluating a program: What is the story behind a successful or unsuccessful program recipient? Can we deliver the same services for less by improving our operating model? Are we targeting the people we said we would target? Are our constituents satisfied with the service we provide? Is there significant demand for the service we provide? Is the demand sustained—do people come back for more? Is the problem we are solving the most pressing in our context?

We could go on. These are the questions that key stakeholders often want answered. Some of these questions can be answered with data. Others are tougher to tackle. But—and this is the crucial point—their answers are not measures of impact.

Alternative: To answer these questions, data collection and analysis need to focus more precisely on the question being asked. Understanding constituent satisfaction requires feedback data. Improving the cost-effectiveness of program delivery requires detailed data on costs by site, as well as by product or service. All of this is important program monitoring data to collect, but none of it requires an impact evaluation.

2. Not Now: The program design is not ready.

Thinking through the theory of change is the first step to planning out a monitoring or evaluation strategy. A theory of change articulates what goes into a program, what gets done, and how the world is expected to change as a result. Without it, staff may hold conflicting or muddled ideas about how or why a program works, which can result in large variations in implementation.

Articulating a clear theory of change is not merely an academic exercise for retreats and donors. A theory of change guides right-fit data collection by making clear what data to track to make sure an organization is doing what it says it does, to provide feedback and engagement data to guide program learning and improvement (neither of which requires a counterfactual), and to provide guidance for key outcomes to track in an impact assessment (which does require a counterfactual to be meaningful).

An untested theory of change likely contains mistaken assumptions. For example, hypothesized connections ("theory") between program elements may not hold. Assumptions may also be wrong empirically: Program outcomes may depend on everyone finishing the training part of the program. Do they? Good management data could help demonstrate this. Similarly, programs may assume that demand exists for their services (e.g., microcredit), but a good needs assessment might show that reasonable credit alternatives exist.

Large impact evaluations undertaken before key assumptions in the theory of change undergo examination are likely to be misguided and ultimately lead to conflict over interpretation. If the program is found not to work, implementers are likely to reject the results, arguing that the program evaluation doesn't reflect current implementation.

Alternative: Validating the initial steps in the theory of change is a critical step before moving on to measuring impact. Consider a program to deliver child development, health, and nutrition information to expectant mothers in order to improve prenatal care and early childhood outcomes. Starting an impact evaluation before knowing if expectant mothers will actually attend the training and adopt the practices makes little sense. First establish that there is a basic take-up of the program and that some immediate behaviors are being adopted. Before starting an impact evaluation of a program providing savings accounts, determine whether people will actually open a savings account when offered, and that they subsequently put money into the account. If not, the savings account design should be reconsidered.

If the theory of change has not been fully developed, then the obvious step is to develop the theory for the program, following the implementation step by step, examining the assumptions being made, and gathering data to test them. Then gather monitoring data on implementation and uptake before proceeding to an impact evaluation. Is the program reaching the people it targets? Are those individuals using the product or service? For how long and how intensively do they use the product or service? Based on this information, how can the program be improved?

When the program is still being adapted and implementation kinks worked out, it is probably too early to evaluate the program's impact. This is a tricky situation. We could craft some general principles for determining when a program is "ready" for evaluation, such as "Basic levels of demand are observed for the program," or "Constituents provide positive feedback." The challenge is then applying these principles to specific situations. Here reasonable people will no doubt disagree, and these principles cannot clearly resolve what to do for any given situation. The most sensible solution is to wait and let the program work out the implementation kinks. If women are not coming to the training or teachers are not following a new curriculum, wait, watch, try new tactics or incentives; and in the meantime, collect good monitoring data that informs progress.

3. Not Now: The program implementation is not ready.

Even if a program's theory has been fully defined and basic assumptions tested, implementation may falter. An evaluation that finds no

impact for a project with weak implementation is hard to interpret. Is the finding the result of poor implementation, the wrong partner, or outside circumstances (e.g., civil unrest or other disturbances)? Either way, when implementation is weak, impact evaluation is a bad choice.

To return to our previous example, a prenatal training program may have determined that mothers demand these services and will show up and complete the training in an "ideal" setting where the program was tested. But what if during program rollout the trainings are not implemented as planned? What if not all mothers complete the training? Basic implementation information is needed before moving to impact evaluation, so that stakeholders are satisfied that the program as designed is (roughly) the same as the program that is implemented. Otherwise, evaluation resources are wasted.

Alternative: Collect good monitoring data and use it to strengthen implementation. Evaluators can either work with program leadership to improve implementation or decide that a certain organization is not a good fit for an impact evaluation.

But what if the real world takes over and politics (or funding) mean you must evaluate now or never? If the program is still not ready, consider again carefully whether impact evaluation is the right step. Will the evaluation help answer theory-based questions under real-world implementation conditions? Will an evaluation now make an innovative or controversial program more likely to be accepted by constituents? Are the technical issues discussed below addressed, and can you construct a reliable comparison group? If you answer no to any of these questions, impact evaluation isn't the right step. But if you answer yes to all, an evaluation of a program that isn't quite ready can still inform important and timely policy-relevant decisions, especially if the evaluators work closely with the policy makers throughout the evaluation process.

4. Not Now: It is too late.

The desire for impact measurement often comes after a program has already expanded and has no plans for further expansion. In these cases, it may be too late. Once a program has begun implementation, it is too late randomly to assign individuals or households or communities to treatment and control. Creating a non-randomized comparison group may be viable but is often hard to do and quite expensive. And the true comparability of this group may still be in question, thus rendering the evaluation less convincing.

Alternative: Plan for future expansions. Will the program be scaled up elsewhere? If so, read on to understand whether measuring impact is feasible. If the program has changed significantly as a result of organizational learning and improvement, timing may be perfect to then assess impact.

5. Not Feasible: Resources are too limited.

Resource limitations can doom the potential for impact evaluation in two ways: The program scale may be too small, or resources may be too scarce to engage in high-quality measurement.

If a program is small, there simply will not be enough data to detect impact unless the impact is massive. Without sounding too sour, few initiatives have truly massive impact. And an impact evaluation with an ambiguous conclusion is worse than doing nothing at all. A lot of money is spent to learn absolutely nothing—money that could have been spent to help more people.

Similarly, if there is not enough money to do a good evaluation, consider not doing it at all. You may be forced to have too small a sample, cut too many corners on what you are measuring, or risk poor implementation of evaluation protocols.

Alternative: If your scale is limited, do not try to force an answer to the impact question. Consider other options. First, perhaps much is already known about the question at hand. What do other evaluations say about it? How applicable is the context under which those studies were done, and how similar is the intervention? Study the literature to see if there is anything that suggests your approach might be effective. If no other evaluations provide helpful insights, track implementation, get regular feedback, and collect other management data that you can use instead.

If money is limited, consider what is driving the cost of your evaluation. Data (especially household surveys) are a key cost driver for an evaluation. The randomization part of a randomized trial is virtually costless. Can you answer key impact questions with cheaper data, perhaps with administrative data? For example, if testing the impact of a savings program, no doubt many will want to know the impact on health and education spending, agricultural and enterprise investment, consumption of temptation goods, and so forth. But in many cases, just seeing increased savings in regulated financial institutions indicates some success.

If that alternative is not viable or satisfactory, then focus on tracking implementation and collecting other management data that you can put to use. Alternatively, of course, you can raise more money. If the knowledge gap on your issue is big enough—you have a widely implemented program that hasn't been tested, for example, or you're trying a new approach in a conflict setting—then funders may be interested in knowing the answer, too.

6. Not Feasible: Indirect effects are difficult to identify, yet critical to the theory of change.

Many programs include indirect effects that are critical to their theory of change. A farming-information intervention, for example, teaches some farmers new techniques and hopes that they share this information with their neighbors and extended family. A health intervention protects individuals from an infectious disease and anticipates that those who come into contact with the treated individuals are also helped, because they will also not contract the disease.

In these cases, a simple question ought to be asked: Does one reasonably believe (and ideally have some evidence from elsewhere) that the indirect effects are significant enough that ignoring them may radically alter the policy implication of the results? If so, then ignoring them could lead to a deeply flawed study—one that should not be done at all.

Measuring such indirect effects correctly is critical to understanding a program's true impact. Take the example of deworming school children. Prior to Edward Miguel and Michael Kremer's 2004 study of deworming in *Econometrica*, studies that tested the impact of school-based deworming typically randomized within schools, with some children receiving deworming pills and others not. Program effects were evaluated by comparing children who received treatment with those who did not. Yet there was good reason to believe that there were indirect effects across children within the same schools—children playing barefoot in the same schoolyard pass infection from one

to the other. So within any given school, the control group also got partially treated. Imagine that this indirect effect is big—so big that it is the same size as the direct effect. Even if treatment had huge effects on health or schooling outcomes, comparing treated and untreated children would lead to the conclusion that deworming has no effect at all. Miguel and Kremer's deworming study explicitly measured these indirect effects. Doing so fundamentally changed the cost-benefit calculation of deworming: With indirect effects included, the benefits of deworming turned out to be quite large.

Alternative: Measuring indirect effects can be a feature of a good impact evaluation, rather than an obstacle. Of course, if indirect effects are ignored, then the presence of such issues can introduce bias, and thus incorrect conclusions.

In considering the response to indirect effects, a first tack is to review existing studies and theory to predict how important these issues are. If they are significant, and therefore important to measure, then there are two potential approaches to take: First, indirect effects can be included in the experimental design—for example, by creating two control groups: one that is exposed indirectly to treatment and the other that is not. Second, data can be collected on indirect effects. Ask participants who they talk to, and measure social networks so that the path of indirect effects can be estimated. If indirect effects can't be accurately estimated, however, and they are likely to be large, then impact evaluation is not a good choice. Resources will be wasted if true impact is masked by indirect effects.

7. Not Feasible: Program setting is too chaotic.

Some situations are not amenable to impact evaluation. Many disaster-relief situations, for example, would be difficult, if not impossible, to evaluate, since implementation is constantly shifting to adapt to evolving circumstances. Maintaining strict experimental protocols could be costly, compromising the quality of the implementation. Even if not costly in theory, such protocols are unlikely to be adhered to in a rapidly changing environment and could prevent assistance from going to those who need it most.

Alternative: Track implementation activities and collect other management data that you can use to strengthen the program. Consider also whether there are operational questions that could generate useful learning. Operational (sometimes called rapid-cycle or rapid-fire or A/B) experiments can help improve implementation: Will sending a text message to remind someone to do something influence short-run behavior? How frequently should that text message be sent, at what time of day, and what exactly should it say? Is transferring funds via cash or mobile money more effective for getting money to those affected? How will lump-sum versus spread-out transfers influence short-run investment choices? Such short-run operational questions may be amenable to evaluation.

8. Not Feasible: Implementation happens at too high a level.

Consider monetary or trade policy. Such reforms typically occur for an entire country. Randomizing policy at the country level would be infeasible and ridiculous. Policies implemented at lower levels—say counties or cities—might work for randomization if there are a sufficient number of cities and spillover effects are not a big issue. Similarly, advocacy campaigns are often targeted at a high level (countries, provinces, or regions) and may not be easily amenable to impact evaluation.

Alternative: A clear theory of intended policy change is critical. Then track implementation, feedback, and management data on whether the changes implied by theory are occurring as expected.

9. Not Worth It: We already know the answer.

In some cases, the answer about whether a program works might already be known from another study, or set of studies. In that case, little will be learned from another impact evaluation. But sometimes donors or boards push for this unnecessary work to check their investments. And organizations may not be sure if the existing evidence is sufficient, leading them to invest in unnecessary impact evaluations "just to be sure."

Alternative: Resist demands for impact measurement and find good arguments for why available evidence applies to your work. In "The Generalizability Puzzle," their Summer 2017 article for *Stanford Social Innovation Review*, Mary Ann Bates and Rachel Glennerster provide some guidance. In short, two main conditions are key to assessing the applicability of existing studies. First, the theory behind the evaluated program must be similar to your program—in other words, the program relies on the same individual, biological, or social mechanism. Second, the contextual features that matter for the program should be relatively clear and similar to the context of your work.

We also suggest that donors consider the more critical issue for scaling up effective solutions: implementation. Use monitoring tools to ask: Does the implementation follow what is known about the program model? Again, track the activities and feedback to know whether the implementation adheres to the evidence from elsewhere. A good example of this is the Catch Up program in Zambia, where the Ministry of General Education is scaling up the proven Teaching at the Right Level (TaRL) approach pioneered by the Indian NGO Pratham. With support from IPA and the Abdul Latif Jameel Poverty Action Lab (J-PAL), teams in Zambia are taking the TaRL program, mapping evidence to the Zambian context, supporting pilot implementation, and monitoring and assessing viability for scale-up.

10. Not Worth It: No generalized knowledge gain.

An impact evaluation should help determine *why* something works, not merely *whether* it works. Impact evaluations should not be undertaken if they will provide no generalizable knowledge on the "why" question—that is, if they are useful only to the implementing organization and only for that given implementation. This rule applies to programs with little possibility of scale, perhaps because the beneficiaries of a particular program are highly specialized or unusual, or because the program is rare and unlikely to be replicated or scaled. If evaluations have only a one-shot use, they are almost always not worth the cost.

Alternative: If a program is unlikely to run again or has little potential for scale-up or replication, the best course of action is to measure implementation to make sure the program is running as intended. If some idea about the "why" is needed, a clear program theory and good implementation data (including data on early outcomes) can also help shed light on why something works. But an investment in measuring impact in this situation is misplaced.

COLLECTING THE RIGHT DATA

As should now be clear, the allure of measuring impact distracts from the more prosaic but crucial tasks of monitoring implementation and improving programs. Even the best idea will not have an impact

if implemented poorly. And impact evaluation should not proceed without solid data on implementation. Too often, monitoring data are undervalued because they lack connection to critical organizational decisions and thus do not help organizations learn and iterate. When data are collected and then not used internally, monitoring is wasted overhead that doesn't contribute to organizational goals.

External demands for impact undervalue information on implementation because such data often remain unconnected to a theory of change showing how programs create impact. Without that connection, donors and boards overlook the usefulness of implementation data. Right-fit systems generate data that show progress toward impact for donors and provide decision makers with actionable information for improvement. These systems are just as important as proving impact.

How can organizations develop such right-fit monitoring systems? In *The Goldilocks Challenge*, we develop what we call the CART principles—four rules to help organizations seeking to build these systems. CART stands for data that are Credible, Actionable, Responsible, and Transportable.

Credible: Collect high-quality data and analyze them accurately.

Credible data are valid, reliable, and appropriately analyzed. Valid data accurately capture the core concept that one is seeking to measure. While this may sound obvious, collecting valid data can be tricky.

Seemingly straightforward concepts such as schooling or medical care may be measured in quite different ways in different settings. Consider trying to measure health-seeking behavior: Should people be asked about visits to the doctor? A nurse? A traditional healer? How the question is framed affects the answer you get.

Credible data are also reliable. Reliability requires consistency; the data collection procedure should capture data in a consistent way. An unreliable scale produces a different weight every time one steps on it; a reliable one does not.

The final component of the *credible* principle is appropriate analysis. Credible data analysis requires understanding when to measure impact—and, just as important, when not to measure it. Even high-quality data to measure impact without a counterfactual can produce incorrect estimates of impact.

Actionable: Collect data you can commit to use.

Even the most credible data are useless if they end up sitting on a shelf or in a data file, never to be used to help improve programming. The pressure to appear "data-driven" often leads organizations to collect more data than anyone can be reasonably expected to use. In theory, more information seems better, but in reality, when organizations collect more data than they can possibly use, they struggle to identify the information that will actually help them make decisions.

The actionable principle aims to solve this problem by calling on organizations to collect only data they will use. Organizations should ask three questions of every piece of data that they want to collect: (1) Is there a specific action that we will take based on the findings? (2) Do we have the resources necessary to implement that action? (3) Do we have the commitment required to take that action?

Responsible: Ensure that the benefits of data collection outweigh the costs.

The increasing ease of data collection can lull organizations into a "more is better" mentality. Weighing the full costs of data collection against the benefits avoids this trap. Cost includes the obvious direct costs of data collection but also includes the opportunity costs, since any money and time spent collecting data could have been used elsewhere. This foregone "opportunity" is a real cost. Costs to respondents—those providing the data—are significant but often overlooked. Responsible data collection also requires minimizing risks to these constituents through transparent processes, protection of individuals' sensitive information, and proper research protocols.

While collecting data has real costs, the benefits must also be considered. We incur a large social cost by collecting too little data. A lack of data about program implementation could hide flaws that are weakening a program. And without the ability to identify a problem in the first place, it cannot be fixed. Too little data can also lead to inefficient programs persisting, and thus money wasted. And too little data can also mean that donors do not know whether their money is being used effectively. That money could be spent on programs with a greater commitment to learning and improvement, or those with demonstrated impact.

Transportable: Collect data that generate knowledge for other programs.

Valuable lessons generated from monitoring and evaluations should help build more effective programs. To be transportable, monitoring and evaluation data should be placed in a generalizable context or theory—they should address the question of why something works. Such theories need not always be complex, but they should be detailed enough to guide data collection and identify the conditions under which the results are likely to hold. Clarifying the theory underlying the program is also critical to understanding whether and when to measure impact, as we have argued.

Transportability also requires transparency—organizations must be willing to share their findings. Monitoring and evaluation data based on a clear theory and made available to others support another key element of transportability: replication. Clear theory and monitoring data provide critical information about what should be replicated. Undertaking a program in another context provides powerful policy information about when and where a given intervention will work. A lack of transparency has real social costs. Without transparency, other organizations cannot identify the lessons for their own programs.

CREATING A RIGHT-FIT SYSTEM

CART provides organizations with a set of principles to guide them in deciding which credible data are most critical to collect. But organizations need to do more than simply collect the right data. They need to integrate the data fully into what they do. They need to develop right-fit evidence systems.

Creating such systems should be a priority for all organizations. First, many organizations will be better served by improving their systems for monitoring and managing performance, rather than focusing on measuring impact. Right-fit evidence systems provide credible and actionable data that are far more valuable than the results of a poorly run impact evaluation. Second, society is better served when organizations develop right-fit evidence systems. High-quality management data help organizations learn and improve. Transparent data that are connected to theory help build our generalized knowledge of what

works—and in what settings. Good programs can be replicated, poor ones retired. Impact evaluations are undertaken only when the conditions are right—avoiding waste and maximizing scarce resources.

The first step in moving toward right-fit evidence happens at the organizational level. To support program learning and improvement, evidence must be actionable—that is, it must be incorporated into organizational decision-making processes. An actionable system of data management does three things: collect the right data, report the data in useful formats in a timely fashion, and create organizational capacity and commitment to using data.

Organizations should collect five types of monitoring data. Two of these—*financial* and *activity (implementation) tracking*—are already collected by many organizations to help them demonstrate accountability by tracking program implementation and its costs. The other three—*targeting*, *engagement*, and *feedback*—are less commonly collected but are critical for program improvement.

The key to right-sized monitoring data is finding a balance between external accountability requirements and internal management needs. Consider *financial* data first. External accountability requirements often focus on revenues and expenses at the administrative and programmatic levels. To move beyond accountability to learning, organizations need to connect cost and revenue data directly to ongoing operations. This way they can assess the relative costs of services across programs and program sites.

Many organizations also collect monitoring data about *program implementation*, including outputs delivered (e.g., trainings completed). But such data are not clearly connected to a decision-making system based on a clear theory for the program. A clear and detailed theory of change supports organizations in pinpointing the key outputs of each program activity so that they can develop credible measures for them.

Targeting data answer the question: Who is actually participating in the program? They help organizations understand if they are reaching their target populations and identify changes (to outreach efforts or program design, for example) that can be undertaken if they are not. To be useful, targeting data must be collected and reviewed regularly, so that corrective changes can be made in a timely manner.

Engagement data answer the question: Beyond showing up, are people using the program? Once organizations have collected activity tracking data and feel confident that a program is being well delivered, the next step is to understand whether the program works as intended from the participant perspective. Engagement data provide important information on program quality. How did participants interact with the product or service? How passionate were they? Did they take advantage of all the benefits they were offered?

Feedback data answer the question: What do people have to say about your program? Feedback data give information about its strengths and weaknesses from participants' perspectives. When engagement data reveal low participation, feedback data can provide information on why. Low engagement may signal that more feedback is needed from intended beneficiaries in order to improve program delivery.

EMPOWERING DATA

Another fundamental challenge to creating an actionable data system is empowering decision makers to use the data to make decisions. Empowerment requires capacity and commitment. Building organizational commitment requires sharing data internally, holding staff members responsible for reporting on data, and creating a culture of learning and inquiry.

To do this, organizations first need the capacity to share the data they collect. This does not require big investments in technology. It can be as simple as a chalkboard or as fancy as a computerized data dashboard, but the goal should be to find the simplest possible system that allows everyone access to the data in a timely fashion.

Next, the organization needs a procedure for reviewing data that can be integrated into program operations and organizational routines. Again, this need not be complex. Data can be presented and discussed at a weekly or monthly staff meeting. The important thing is that data are reviewed on a regular basis in a venue that involves both program managers and staff.

But just holding meetings will not be enough to create organizational commitment and build capacity if accountability and learning are not built into the process. Program staff should be responsible for reporting the data, sharing what is working well, and developing strategies to improve performance when things are not. Managers can demonstrate organizational commitment by engaging in meetings and listening to program staff. Accountability efforts should focus on the ability of staff to understand, explain, and develop responses to data—in other words, focus on learning and improvement, not on punishment.

The final element of an actionable system is consistent follow-up. Organizations must return to the data and actually use it to inform program decisions. Without consistent follow-up, staff will quickly learn that data collection doesn't really matter and will stop investing in the credibility of the data.

To simplify the task of improving data collection and analysis, we offer a three-question test that an organization can apply to all monitoring data it collects:

- Can and will the (cost-effectively collected) data help manage the day-to-day operations or design decisions for your program?
- Are the data useful for accountability, to verify that the organization is doing what it said it would do?
- Will your organization commit to using the data and make investments in organizational structures necessary to do so?

If you cannot answer yes to at least one of these questions, then you probably should not be collecting the data.

Maybe this seemingly new turn away from impact evaluation is all a part of our plan to make rigorous evaluations even more useful to decision makers at the right time. When organizations or programs aren't ready for an impact evaluation, they still need good data to make decisions or improve the implementation of their model. And when a randomized evaluation (or six) shows that something works and it is ready for scale, a good monitoring system based on a sound theory of change is the critical link to ensuring quality implementation of the program as it scales.

In the interim, our plan is to shift the focus to evidence strategies that build learning and improvement. If this stratagem ultimately leads to more effective impact evaluations, so much the better. ■