We at the Institute of Philanthropy have learned many lessons about measuring social impact.

This supplement collects some of the best research on impact measurement published in STANFORD SOCIAL INNOVATION REVIEW to further discussion, collaboration, and co-learning opportunities about charitable giving in Asia and beyond.

THE BEST OF SSIR:
IMPACTFUL PHILANTHROPY IN THE REAL WORLD

SUPPLEMENT ILLUSTRATIONS BY DAVID PLUNKERT

IMPORTANCE OF IMPACT MEASUREMENT

THEORY: IMPACT MEASUREMENT FRAMEWORKS

PRACTICE: OPERATIONALIZATION OF IMPACT MEASUREMENT

A PLATFORM FOR SHARING KNOWLEDGE

In exploring how to measure impact, we seek partners to share insights and develop frameworks to help solve social problems.

BY WINFRIED ENGELBRECHT-BRESGES

The Institute of Philanthropy (IoP) is excited to present this special supplement in partnership with *Stanford Social Innovation Review* (*SSIR*). We are eager to discover how philanthropy can effect real change when it comes to pressing social issues. We also see opportunities for Asian perspectives to contribute to the global debate on how exactly to go about measuring impact, including the insights featured in the pages ahead.

Philanthropy plays a special role in the betterment of society. Its flexibility allows for intervention and experimentation in line with—and often ahead of—government policy and private sector provisions. As governments face sweeping challenges like climate change and aging populations and grapple with competing priorities, philanthropy is becoming more important. This raises the stakes for the sector to demonstrate its impact and the

extent to which that impact lasts. (See "In Search of Durable Change" in this supplement.)

We can trace the concept of social impact back to the 19th century, but it did not enter the mainstream until the second half of the 20th. This was due to two converging forces: a desire among philanthropists to make a difference and the notion of corporate social responsibility. Globally, philanthropy's focus has shifted from measuring the amount of charitable giving to measuring how it creates positive change in society ("A Playbook for Designing Social Impact Measurement"). Companies and investors now embrace a more integrated approach to generating value that encompasses economic value, social impact, and environmental issues.

The advent of an environmental, social, and governance (ESG) framework was a significant step toward increasing the transparency required to drive change. And yet, the difficulty of quantifying social impact, for example in estimating reductions in greenhouse gas emissions, has blurred lines of accountability and undermined the *S* in ESG reporting. This has damaged the framework's credibility in the commercial sector ("Fixing the *S* in ESG").

The United Nations' Sustainable Development Goals may represent the most successful effort to align stakeholders with a uniform approach to social impact. But the UN designed them to correspond to its global development agenda. They spell out high-level priorities such as poverty reduction and mortality rates, but they are not easy to translate into accounting for philanthropic dollars. In other words, they don't provide a standardized yet widely appropriate approach to measuring social impact ("Prioritizing Impact Measurement in the Funding of Social Innovation").

In Asia, home to some of the most populous nations and fastest-developing economies, we are experiencing rapid growth in philanthropic giving commensurate with the region's rising wealth. As Asian foundations work to professionalize philanthropic practices, it is imperative to clarify how we approach, interpret, and use social impact measurement ("Putting Evidence to Use").

It is against this backdrop that the IoP was established in September 2023 with a seed donation of HK$5 billion ($640 million) from The Hong Kong Jockey Club and its Charities Trust. We are an Asia-based "think-fund-do" tank that seeks to bridge regional and global philanthropic thinking for the betterment of societies everywhere.

We aspire to integrate Asian experiences and insights into the global philanthropy ecosystem by bringing together thought leaders, philanthropists, and stakeholders from around the world to exchange best practices and learn from one another. We also aim to cocreate and cofund philanthropic projects. These include collaborations with The Rockefeller Foundation to scale equitable solutions addressing the impacts of climate change and with China's National Health Commission to build capacity in its health-care sector to strengthen post-pandemic preparedness.

We present this *SSIR* supplement to consolidate and share knowledge on social impact measurement that can help philanthropy professionals leapfrog barriers, especially in Asia. The articles that follow highlight different perspectives on measuring impact and present approaches that are multifaceted in theory and practice, such as the appropriate use of impact evaluation ("Ten Reasons Not to Measure Impact") and the importance of stakeholder feedback in impact assessment ("Time for a Three-Legged Measurement Stool"). They also include inspirational and enlightening real-life lessons from organizations such as the Abdul Latif Jameel Poverty Action Lab (J-PAL) and its evidence-informed innovations ("Beyond Randomized Controlled Trials"); Rohini Nilekani Philanthropies and its approach that extends far beyond simple metrics ("Plotting Impact Beyond Simple Metrics"); and The Hong Kong Jockey Club Charities Trust ("Adapt and Innovate").

Many Asian foundations are just starting out on their impact measurement journeys and working through the challenges of devising a framework to measure social impact in a way that supports their goals. By sharing our lessons, we hope to help peers to fast-track their own discovery process and avoid dead ends.

We also hope that our knowledge-sharing platform can contribute to more meaningful dialogue, actionable collaboration, and co-learning opportunities that will help shape the future of charitable giving in Asia and beyond.

Winfried Engelbrecht-Bresges is a director of the Institute of Philanthropy and CEO of The Hong Kong Jockey Club.

ADAPT AND INNOVATE

Impact measurement evolves with changing times and circumstances. That dynamic offers opportunities to innovate, as the HKJC Charities Trust found.

BY GABRIEL M. LEUNG

Impact measurement is a useful indicator of how philanthropy can make a difference. For an organization with a broad remit, such as The Hong Kong Jockey Club Charities Trust ("the Trust"), which serves Hong Kong's changing needs, finding the right metrics to assess our collective social impact is challenging. We recently developed a new approach that measures the collective impact of individual funded projects, according to different funding themes, and across the entire organization. Sharing some lessons with our regional and global peers, we hope to contribute to the evolution of impact measurement and better serve our grantees and, through them, our beneficiaries.

THE EVOLUTION OF IMPACT MEASUREMENT AT HKJC CHARITIES TRUST

The Trust's primary goal is to contribute to the betterment of society, which has meant different things at different moments in our century-long history. A key part of our evolution has been adapting both our funding goals and our accountability to address changing societal needs.

Since our first recorded charitable donation in 1915, contributions have continued to climb, especially starting in the 1950s, when Hong Kong's

The Hong Kong Jockey Club Charities Trust, among the world's largest philanthropic foundations in terms of annual giving, aims to address community needs for the betterment of Hong Kong. In its 2022/23 fiscal year, its total approved donations amounted to $935 million, benefiting 247 charity and community projects across six issue-based program areas.

population doubled in 20 years, from two million in 1951 to four million in 1971, mostly due to migration from mainland China.

With rapid population growth putting acute stress on government finances, the Trust came forward to collaborate with local authorities to fund projects such as building schools and hospitals, and then to expand this work by constructing arts facilities, sports grounds, public parks, and other infrastructure to meet the changing needs of a young, prospering society and its growing middle class. (See "Philanthropic Giving for a Changing Society" below.)

In the early 2000s, the focus of our giving shifted to service programs to tackle emerging issues such as an aging population and environmental protection. Since each program had different goals and required different approaches, we promoted pledged outputs and clear targets for each project to ensure prudent funding allocation, accountability, and transparency. We also began developing tailored, rigorous impact measurement indicators for large-scale university-led programs. These measures began as the exception rather than the rule, but by the 2010s, our approach changed. We recognized that applying these rigorous measures across all programs could optimize impact.

This objective was not straightforward by any means. First, the local nonprofit sector had been unfamiliar with impact measurement and did not know how to implement it or where to find the required resources (funding, methodology, expertise, and manpower). Second, the breadth of our programmatic coverage made it challenging, from the Trust's perspective, to identify a single common metric to assess impact.

After consulting grantees and external experts, we took a "lowest common denominator" approach by adopting a rudimentary measure for project impact based on changes in the knowledge, attitudes, and behaviors of beneficiaries. This simple and potentially universal framework could be applied to all projects and allowed grantees to adapt their execution according to their circumstances. We provided training and a basic template to help them produce indicators and survey questions.

The framework's simplicity turned out to be a limitation. It lacked scientific rigor, and there was little evidence to show that shifts in knowledge and attitudes led to sustained behavioral changes.[1] As the nonprofit sector became stronger and more professional, we believed that a more scientifically valid approach would better maximize value-added impact for every dollar invested and align our work with global trends.

This realization dovetailed with our organization's evolving agenda. We wanted to understand the collective impact of funded projects within a specific field, not only the impact of each individual project. Two years ago, we launched six program areas dedicated to different issues: Positive Aging and Elderly Care; Youth Development and Poverty Alleviation; Healthy Community; Talent and Sector Development; Sports and Culture; and Sustainability. We are also testing a novel approach to assess impact that aggregates across all fields and levels of giving.

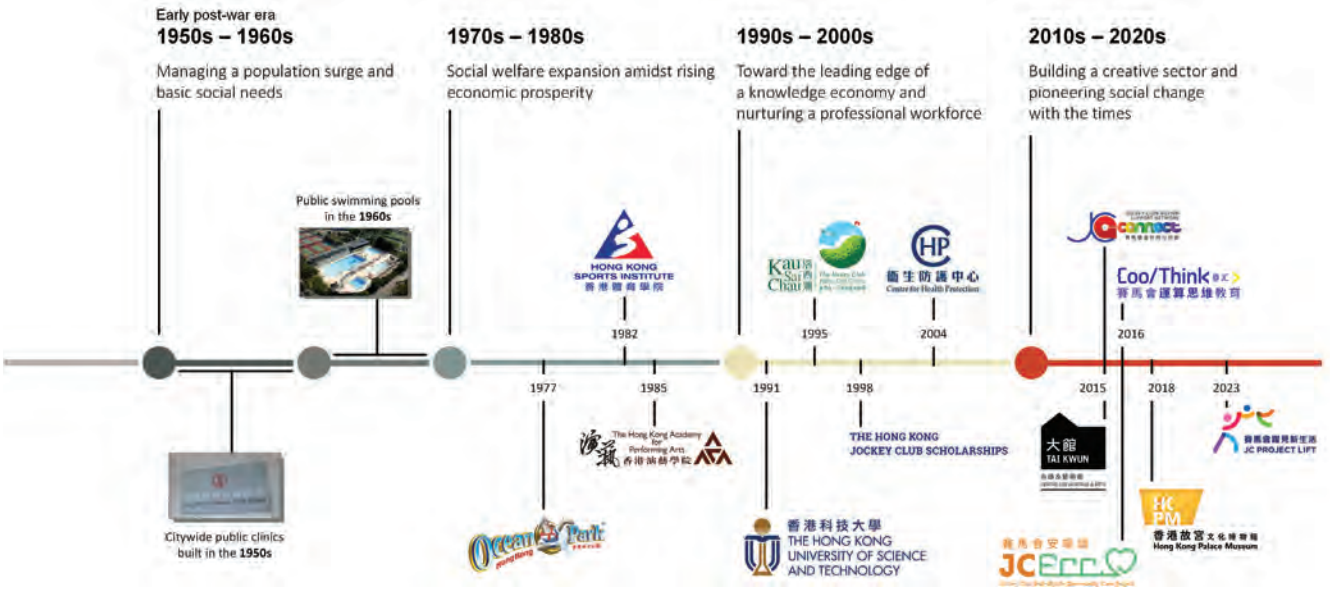
A THREE-TIERED, HIERARCHICAL APPROACH

Following a comprehensive literature review of studies and practices, our latest approach focuses on a hierarchical model that considers impact at the project, program, and organization (HKJC Charities Trust) levels. (See "The Hierarchical Model" on page 7.) The aim is to produce timely, relevant, and rigorous evidence based on the following core principles:

1. clear objectives and hypotheses to define the purpose and scope of impact measurement;
2. rigorous evaluative designs to establish causality between interventions and outcomes;
3. validated measures to provide accurate and appropriate outcome data;
4. robust statistical methods to quantify the effects attributable to the interventions;
5. aggregation of impact to demonstrate social impact at multiple levels; and
6. regular, fit-for-purpose reporting.

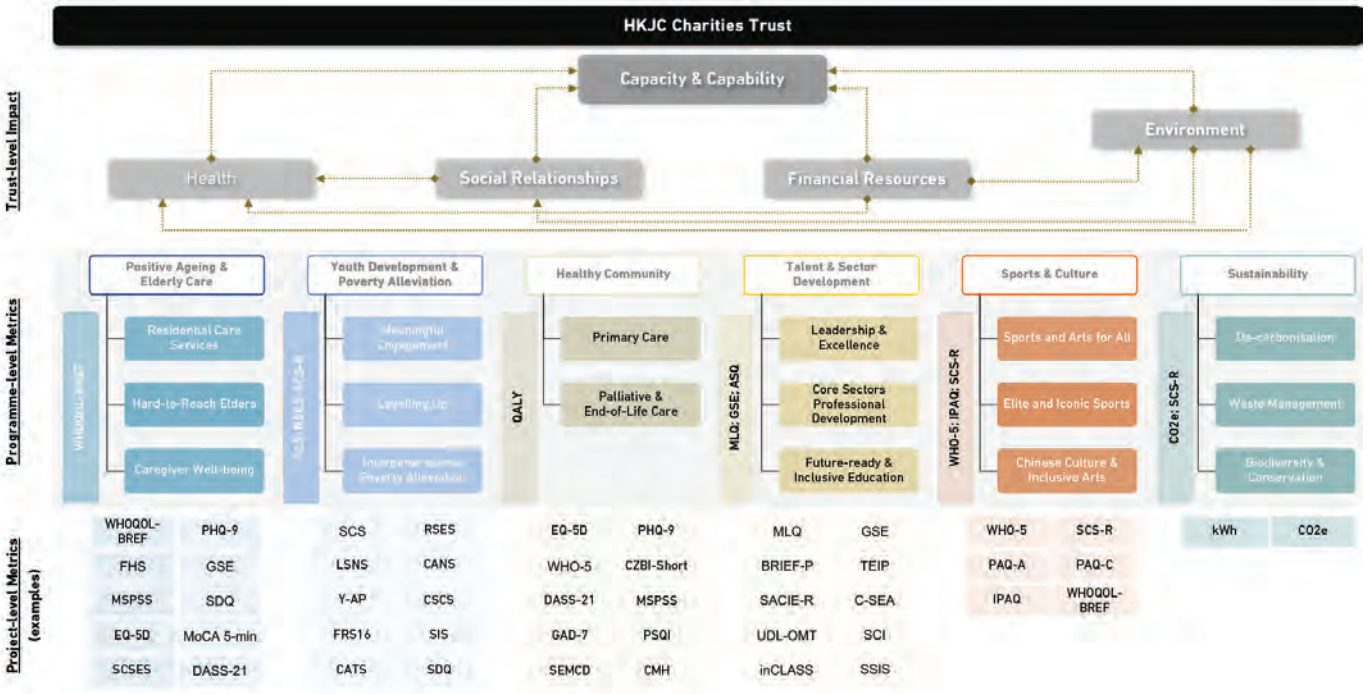
Philanthropic Giving for a Changing Society

THE TIMELINE PLACES HKJC CHARITIES TRUST GIVINGS IN POSTWAR ERAS OF SOCIO-ECONOMIC DEVELOPMENT IN HONG KONG.



The Hierarchical Model

THE HKJC CHARITIES TRUST CONSIDERS IMPACT AT THE PROJECT, PROGRAM, AND ORGANIZATION LEVELS.



At the project level, we apply many different, appropriate measurement instruments, such as the abbreviated version of the World Health Organization's Quality of Life assessment (WHOQOL-BREF) for tracking positive aging and elderly care and the International Physical Activity Questionnaire (IPAQ) for tracking sports participation. For a glossary of all metrics, see: ssir.org/articles/entry/impact-measurement-hkjc.

At the **project level**, each project should employ a robust evaluation design, such as a randomized controlled trial or longitudinal cohort study, supported by validated measurement indicators and appropriate analytic methods that aim to establish causality. This methodology helps us to monitor and evaluate projects and to provide useful feedback to funding applicants.

At the **program level**, thematic indicators are applied across multiple projects so that the combined impact on a particular problem or theme can be assessed. For instance, if multiple elderly care projects provide different services, the collective impact could be determined by using process measures, such as the total number of people served, and the end goals, such as health/disability burden outcomes. Both quantitative and qualitative outcomes are important considerations in creating a comprehensive view of impact that includes all beneficiaries. We believe that defining overarching program-level objectives and their corresponding metrics can help us articulate and report on social benefits generated across different but cognate projects.

At the **Trust (organization) level**, we assess whether and to what extent the Trust fulfills its underlying mission of bettering society. Project indicators can be transformed to dimensionless form ("effect size") and aggregated using statistical techniques such as meta-analysis across five core domains in which projects strive to create positive change: Capacity and Capability; Health; Social Relationships; Financial Resources; and Environment.

It is still early days, and especially at the organization level, we continue to explore how to frame and validate our findings. We can look to our global peers, but we are also committed to finding evidence rooted in the East and in the context of communities at home.

FOUR TAKEAWAYS

Our journey toward measuring and validating our impact is a work in progress, and yet we have already learned several important lessons. We invite fellow philanthropists and practitioners to join us in this conversation.

1. There is no one-size-fits-all framework. Every project takes place in a particular context that must be considered when designing impact measurement. This context can shift, as the Trust experienced when our giving moved from building basic social infrastructure to focusing on service programs and related policy innovations. Additionally, impact measurement means something quite different to a large organization such as the Trust when compared with individual family offices, which may not enjoy as much experience or scale in strategic giving and impact measurement.

Government policy can also shape impact measurement. China's 14th Five-Year Plan, for example, sees philanthropy as a key force for distributing wealth and achieving "common prosperity."[2] In India, a mandatory 2 percent of company net profits help fund corporate social responsibility activities to ensure that businesses adopt "inclusive growth."[3] The Muslim zakat tradition of donating a portion of wealth to charitable causes is another example. Policy enables the scaling up of available resources for doing good, which

has implications for impact measurement approaches as stakeholders seek to ensure that their contribution is deployed in the best way possible.

These cases underscore how impact measurement must be fit for purpose, rather than one size fits all.

2. A user-centric mindset is essential. Users, especially grantees and beneficiaries, are the backbone of impact measurement, providing everything from beneficiary data collection to reporting to funders. Procedures must be designed in collaboration with users without becoming overly complex or duplicating data requirements. Users should not face an undue burden, and their feedback should be facilitated in the process.

But funders should also help users gain an awareness of the many impact measurement tools they might use. Funders should provide resources and training while maintaining dialogue with users to align expectations. Both funders and users need to learn and evolve together to drive sustainable impact.

3. Trust is paramount. How we measure impact can strengthen trust-based relationships between funders and grantees and foster a more collaborative and effective philanthropic partnership. Clear impact measurement and transparency will generate accountability in the grantmaking process and a sense of shared responsibility for creating impact. A two-way or multi-way dialogue ensures that grantees have a strong voice in shaping current and future initiatives. This participation can lead to more effective and enduring philanthropic endeavors.

4. Upholding integrity in impact measurement. The Trust's funded projects often target society's most vulnerable groups, whose agency is often ignored and whose trust must be earned. We work with grantees to ensure that beneficiaries choose to participate, that their privacy is respected, and that both data and impact measurement will be responsibly used for their benefit. As advocated by the Stanford Center on Philanthropy and Civil Society, organizations should manage data in line with their mission and with regard to privacy, permission, openness, and pluralism.

LEADING BY ADAPTING

How can funding organizations and their recipients know if they are achieving their objectives? This question must be considered in context. Rigorous, evidence-based impact measurement provides answers, but societal demands and circumstances also affect goals, methods, and applications. The funding organization's goals matter, too. The Hong Kong Jockey Club Charities Trust now strives to articulate our impact at all levels of giving, including at the organization level, across all fields of philanthropic investment. This path has been less trodden, but we feel the stones along the way and continue to evolve and adapt toward greater impact for the betterment of the society we serve.

Gabriel M. Leung is executive director for charities and community at The Hong Kong Jockey Club and a director of the Institute of Philanthropy.

NOTES

[1] Colin Jerolmack and Shamus Khan, "Talk Is Cheap: Ethnography and the Attitudinal Fallacy," *Sociological Methods & Research*, vol. 43, no. 2, 2014; Geoffrey Evans and John Durant, "The Relationship Between Knowledge and Attitudes in the Public Understanding of Science in Britain," *Public Understanding of Science*, vol. 4, no. 1, 1995.

[2] Section 48(3) of the 14th Five-Year Plan for Economic and Social Development and Long-range Objectives Through the Year 2035 of the People's Republic of China.

[3] Section 135(5) of the Companies Act in India, 2013.

A PLAYBOOK FOR DESIGNING SOCIAL IMPACT MEASUREMENT

Thinking about social impact measurement on a spectrum can help organizations develop a clear, evidence-based idea of how or why their programs work.

BY GWENDOLYN REYNOLDS, LISA C. COX, NICHOLAS FRITZ, DANIEL HADLEY & JONATHAN R. ZADRA

Would you buy something from Amazon if it only had one review? Or go to a restaurant that had just five reviews on Yelp? Maybe, but for many of us, it would feel like a risk. The more reviews, the more confident we tend to feel about the quality of a product or place. That's because one review is an anecdote, but 50, 100, even 1,000 reviews is *data.*

If we decide what to buy and where to eat based on data, we should certainly use data to decide where to put resources toward solving social problems. But of course while using data to measure the social impact of a program sounds straightforward, if we misread data or give one data point too much weight, we can end up throwing money away on efforts that don't create real change. Consider a nonprofit working with the local government to end homelessness in a community. If the nonprofit focuses only on individual stories and doesn't measure the number of individuals it serves over time, the local government will never know whether the demand for homelessness services is increasing. Even worse, it won't have the data to know if homelessness in the community is improving or intensifying, which can derail effective resource planning.

Making real social progress means using the right data—and lots of it—to evaluate outcomes, but caveats and misunderstandings abound, even among professionals in the impact measurement arena. Many organizations simply don't have a clear, evidence-based idea of how or why their programs work, and different organizations have different ideas of what impact measurement entails.

As basic as it might sound, one of the most important elements to understand about claims of social impact is the old adage "correlation doesn't equal causation." While correlation, which is simply a relationship between two things, can be a useful endpoint, it's important to distinguish between a lightly informed decision and an evidence-based one, especially when vulnerable populations and billions of dollars hang in the balance.

THE SPECTRUM OF IMPACT MEASUREMENT

Before making important decisions about allocating resources, organizations need to first identify where a program is, where they want it to be, and how to get it there.

To aid this process, we developed a Spectrum of Impact Measurement tool with two axes—level of confidence and level of difficulty—and five stages of impact assessment. (See "Spectrum of Impact Measurement" on page 9.) While the first stage of assessment (determining a theory of change) may be significantly easier than the most advanced stage

(conducting a randomized controlled trial, or RCT), on its own, a theory of change won't provide a high level of confidence in measurement practice. Meanwhile, an organization that is just developing a theory of change will have a much harder time implementing an RCT than an organization that has already identified its key performance indicators (KPIs) and is well on its way to collecting data on them. By breaking down the process into five sequential goals, organizations can take a more "bite-sized" approach to advancing their assessment.

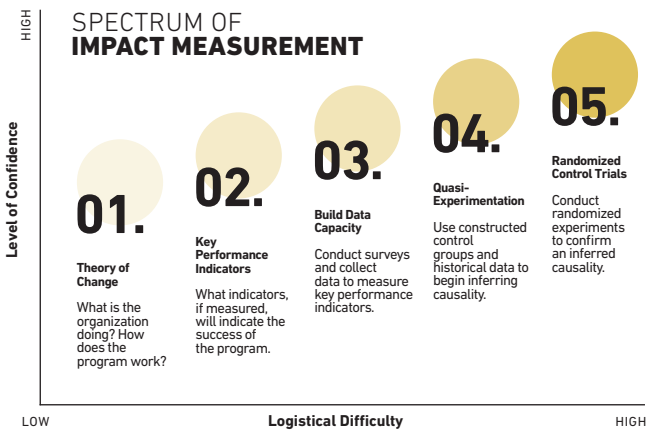
It's worth noting that there is typically a good reason why an organization's program assessment is at a particular level, and if it's on the lower end, it usually isn't because the organization lacks interest in proving its program works. Moving up the spectrum can be time- and resource-intensive, and many organizations have a limited amount of both.

It's also worth noting that achieving the experimental methodologies at the very top of the spectrum need not always be the goal. If an organization isn't pursuing a high level of confidence in its impact, it can gain many insights by simply tracking internal data. A Head Start provider whose goal is to monitor the developmental progress of its children so that it can identify best practices at the classroom level doesn't need a rigorous study with control groups. However, a Head Start provider pursuing a pay-for-performance contract, where it will get paid only if it can prove that its intervention *caused* child outcomes, requires a much higher degree of confidence.

THE FIVE STAGES OF IMPACT MEASUREMENT

The first three points on the spectrum—logic model, KPI selection, and data collection and analysis—are the social sector equivalents of business analytics. They are not simply prerequisites to experimental evaluations; they are valuable in their own right. Developing these three areas helps organizations build cultures that emphasize the importance of data and information, make informed resource allocation decisions, drive performance through goal-setting frameworks and feedback loops, and ultimately use the information they produce as a strategic asset.

The last two steps are evaluations and involve constructing a control group. Only when there is a control group, or a group similar to existing clients that doesn't participate in a program, can organizations begin claiming causal inference—that is, confidently claiming that the program is responsible for the change in clients' circumstance, not merely correlated with it.



Thinking about measurement on a spectrum can help organizations understand where their current measurement practices stand and outline next steps. (Image courtesy of Sorenson Impact Center)

Step 1:
Theory of Change

A theory of change, or logic model, is the foundation for what an organization does and why it does it. It should answer:

1. What impact do you hope to achieve?
2. What is the mechanism by which you achieve that impact?
3. How will you know when you've achieved it?

The second question is where many organizations stumble. Focusing on exactly how a program works, or its "active ingredients," helps inform later stages on the spectrum. The desired impact of an organization focused on afterschool programs, for example, may be improved school grades, but it may struggle with defining what it's about the program that leads to better grades. Improved grades might be due to more hours of education, a reduction in exposure to negative home environments, or even parents providing more or better food because they can now work in the afternoons and make higher wages.

To build a strong theory of change, organizations must think through each possible mechanism and consider how their programs *intend* to achieve impact. Only then can they define appropriate KPIs, set goals, collect data, provide program feedback, and ultimately make decisions more effectively. A theory of change also helps organizations communicate effectively with stakeholders about what their programs are and how they work.

Step 2:
Key Performance Indicators

Many social program providers have a requisite set of metrics to track, usually dictated by funders, and sometimes it can be difficult to focus on the important ones. Organizations should determine which metrics actually matter based on their theory of change, focusing on those that are the broadest reaching, provide the most insight into program implementation, and define success, and remembering that sometimes less is more.

Some metrics can be difficult to measure, because they require data from outside an organization. A Head Start provider, for example, may find it challenging to gather data on student test scores once the children are in kindergarten. However, once an organization establishes its "wishlist metrics," it can identify which of them it already collects and which are within the bounds of possibility. Some readily available metrics will be stepping stones or proxies to wishlist metrics, including simple "counting" metrics, such as the number of people a program serves or client demographics. If Head Start wants to measure kindergarten readiness but can't get school-district data, a proxy measure would be to assess students' kindergarten readiness before and after their year in Head Start.

It should be clear who will be responsible for the collection and analysis of data for each metric, what data they need to measure that metric, and where they will store that data. Each metric should also have an associated goal or goals against which the organization can measure progress over time. A workforce development provider's goal, for example, is to have a certain percentage of enrollees complete its program, and then track progress of that goal. The program should make progress visible to internal stakeholders on a continuous basis and formally review it at set intervals to inform programming decisions. A transparent process demonstrates to frontline staff that the data they're spending precious time collecting

and analyzing is important, and thus helps improve data quality and build a culture that values information and evidence.

Step 3:
Data Collection and Analysis

Organizations often have to report on their programs using a specific data collection system dictated by funders or other partners that doesn't allow them to review the individual-level data they enter into it. For example, some states require that early intervention providers enter client-level survey data, but then do not allow those providers to access it. The state gives the providers an overall score, but doesn't offer enough detail for the provider to know whether or not they are improving their survey scores for specific populations. In these cases, organizations should set up an alternative system that collects the data they need.

In terms of collecting external data, surveys are the most common method, and a well-designed survey will yield accurate, valid, and reliable information. It can be difficult to design a well-thought-through survey, but depending on the issue area, organizations can look for well-validated ones by searching Google Scholar.

Finally, how organizations use and analyze the data they collect is paramount to improving programs. Collecting time series data (the same data point collected at different points in time) allows them to examine how programs change over time. If a Head Start's kindergarten readiness scores are improving every year, for example, it indicates that the program outcome is moving in the right direction. Demographic data and survey results can provide insights into which groups benefit from a program over others. If a homelessness provider finds it's primarily providing services to single individuals rather than families, it can then determine whether to focus on individuals or attempt to adjust to better serve families. Correlations are a powerful sign that a program is either moving outcomes in the right direction or that it's not having the intended effects and needs changing.

Step 4.
Quasi-Experimental Design

Many organizations will be comfortable stopping with correlational data, but without a control group, it's often hard for an organization to prove that its program is responsible for a given change in outcomes.

When organizations want to claim causal inference using a control group but can't randomly assign who receives an intervention and who doesn't, they conduct constructed experiments. Random assignment may be impossible or unethical for many reasons, including in cases where one of the assignments will likely produce a better outcome for the participant. We would never deny hungry families food to create a rigorous experiment. Rather than use random assignment, organizations can find a group of people similar to their clients in several ways.

One option is for an organization to use its own clients as a control group by collecting historical data on the same people as the control. For example, if a workforce development provider wants to track the impact its program had on their clients' earnings, it could collect earnings data for its clients for the year before they entered the program and compare that to their clients' earnings the year after they completed the workforce development program. This has many benefits and works for quite a few programs. However, if there are other potential explanations for changes over time, such as changing economic conditions that may impact wages, disambiguating these external factors from the effect of the program is impossible.

Another option is to find a similar group of people who were not able to participate in a program. For example, if 500 people applied for a housing voucher, but an organization had only 100 housing vouchers to distribute, it could collect data on both those who received housing vouchers and those who didn't, and compare their outcomes. (Note that when resources are limited, the ethical question of not providing services to all who wish to receive them is moot.)

Organizations can also make use of data they already collected for another purpose. A new preschool program, for example, could arrange to receive K-12 data from the school district where it operates. It could then compare the academic outcomes of its preschoolers to the preschoolers in the school district as a whole, while controlling for as many variables as possible.

There are some drawbacks and limitations to the impact an organization can claim when it can't conduct an RCT. The multitude of considerations is difficult to fully address, and therefore we cannot be as confident that a program is the reason for the change in client outcomes. But done correctly, these methods add weight and confidence to impact measurement over more simplistic, correlational analyses.

Step 5.
Randomized Controlled Trial

RCTs have always been considered the gold standard when trying to determine what works. Scientific research has used them for more than 65 years, and they lend organizations the most confidence in their impact. We know that antibiotics treat bacterial infections, for example, because of RCTs conducted in the 20th century.

More recently, it has become best practice to evaluate social programs using RCTs that can be replicated. Multiple RCTs, for example, have shown that the nurse-based home visitation program Nurse-Family Partnership (NFP) reduces child abuse and neglect, and improves cognitive and behavioral outcomes for children. As a result, more local governments and states are funding NFP programs in their own communities.

The keys to designing an RCT are: (1) Each participant has an equal probability of being in either the experimental or control group, and (2) the participants are assigned randomly. So while there may be differences between individuals in each condition, these differences should be randomly distributed and not affect the groups differently. It's important that the program otherwise treats groups similarly throughout the study period, and that it tracks them based on the same metrics and at the same times to allow comparison.

RCTs are designed to compensate for what some call "the fundamental problem of causal inference." Put simply, one cannot both do something in the world *and* observe what would happen if they did nothing. The method is the closest we can get to a time machine, allowing organizations to implement a program *and* estimate what might have happened if they did not.

One of the largest deterrents to organizations conducting RCTs is cost, especially if a program is in its early stages. When there is a strong desire or need to conduct an RCT, organizations can sometimes partner with an academic institution or professional evaluators to design and manage an RCT from start to finish. Having an outsider conduct the evaluation also assures funders that the conclusions of the evaluation were independently verified.

CLAIMING IMPACT WITH CONFIDENCE

Once an organization is collecting and analyzing data, it can begin to make claims about its impact. The strength of these claims is limited to correlation

when the organization stops at step three, but oftentimes, showing and understanding correlational change is enough to strengthen the organization's internal feedback loops and satisfy funders.

The last two steps allow organizations to make causal claims of a program's impact, and they are much easier to accomplish once the first few steps are well developed. Some social programs have no need to conduct a quasi-experiment or RCT, because their programs have been or are already being studied extensively. (The Rigorous Evaluations initiative of the Laura and John Arnold foundation has compiled a thorough catalogue of these.)

In the social sector, impact measurement has been a catchall phrase that often means using easy-to-access data to make big claims. A lack of common understanding about exactly what impact measurement is and what it entails has left many organizations without a playbook for designing and implementing an impact measurement program. The benefit of understanding the spectrum lies not only with the organizations that provide services, but also with the funders who support them. Furthermore, funders who insist that grantees reach the far end of the spectrum need to provide support, financial and otherwise, to achieve the requested level of confidence.

Gwendolyn Reynolds (@gwendoesdata) is a director at the Sorenson Impact Center. She has a bachelor's degree from the University of Utah and master's degree in theological studies from Harvard University.
Lisa C. Cox (@lcmazz) is the communications manager for the Sorenson Impact Center. She holds a bachelor of science degree from Cornell University and a master's degree in journalism from Harvard University in Extension Studies.
Nicholas Fritz (@NicholasMFritz) is a director at the Sorenson Impact Center. He has a bachelor of science degree from the University of Akron and an MBA from the David Eccles School of Business at the University of Utah.
Daniel Hadley (@danielphadley) is a managing director over data, policy, and performance innovation at the Sorenson Impact Center. He is a graduate of the University of Utah and Harvard University, where he received a master's degree in urban planning.
Jonathan R. Zadra (@JonathanZadra) is a director at the Sorenson Impact Center over data, policy, and performance innovation. He completed master's and doctoral degrees in cognitive psychology at the University of Virginia and a bachelor's degree at San Diego State University.

PRIORITIZING IMPACT MEASUREMENT IN THE FUNDING OF SOCIAL INNOVATION

Impact investors and grant makers can learn from each other.

BY LISA HEHENBERGER

A decade ago, I worked on impact measurement guidelines for what was then called the venture philanthropy and social investment field in Europe. At the European Venture Philanthropy Association, we came up with five steps to measure impact for investors and support social enterprises.[1] One of the things we learned through that work is that it's more important to look at how an organization uses impact data in its management—for what purpose and for whom—than the specific data point that emerges from the analysis.

Since then, initiatives such as the Impact Management Project and the Operating Principles for Impact Management (OPIM) have successfully promoted the integration of data in impact-investing management processes so that investors can learn from it and make adjustments as needed. OPIM requires that investors provide independently verified disclosure statements, for example, to show that they are considering impact throughout the investment process, from deal screening through exit.

On the grantmaking side, from our work with foundations and discussions with people like Jeremy Nicholls, more foundations are acknowledging that positively impacting their target populations requires that they formalize processes for listening to beneficiaries, learning from mistakes, and implementing corrective actions.

While the field has become more sophisticated, it still has a long way to go in terms of transparency and accountability. Impact reporting still mainly displays aggregate output figures without explaining methodology or learnings. This contributes to a tension between the standardization of methodology, which enables the kind of benchmarking financial markets require, and generating opportunities for learning so that grant makers can take corrective actions. Both impact investors and grant makers have an opportunity to understand if they are channeling resources to organizations and projects that are tackling the most important problems and making a difference in the lives of beneficiaries. The two fields also stand to learn from each other.

DEVELOPING IMPACT ACCOUNTING AND SETTING NEW STANDARDS

While international accounting standards governing financial accounts dictate what the accounts should look like and how they should be audited, there aren't yet equivalent standards surrounding impact. This creates the impression that impact is subjective, hinders benchmarking and the use of impact data as a basis for decision-making. The International Sustainability Standards Board represents one positive development on this front. Its nonfinancial reporting standards aim to provide investors and other capital-market actors with rigorous and homogeneous information on sustainability risks and opportunities, and thus enable more informed decision-making. The standards make sustainability issues financially material so that shareholders can assess the potential consequences of sustainability decisions on companies' financial statements.

A parallel development is the Impact Economy Foundation's impact-weighted accounts framework, which incorporates the concept of "double materiality." The framework includes an integrated accounting system that considers both an organization's financial materiality and its effect on its environment and stakeholders[2] (essentially impact), and thus helps position finances and impact as integral components of management and corporate governance.

Meanwhile, the United Nations Development Programme has developed a set of internal decision-making standards that will include an assurance framework and a seal of approval. These Sustainable Development Goals (SDGs) Impact Standards aim to help enterprises, bond issuers and investors, and development finance institutions facilitate decision-making so that they can maximize their contributions toward the SDGs.

The field may also take guidance from Europe, which is promoting environmental sustainability through policies like the European Green Deal. Approved in 2019, this policy includes incentives to promote investment in companies and activities that move Europe toward zero emissions while leaving no one behind. It also includes a taxonomy that classifies activities as sustainable, as well as regulation on sustainability disclosure for companies (Corporate Sustainability Reporting Directive) and financial institutions (Sustainable Finance Disclosure Regulation, or SFDR).

European impact investors have welcomed the adoption of SFDR as a means of catalyzing further investment in funds and financial institutions that are working in sustainable ways. This regulation builds on the concept of double materiality, and takes into consideration how investors integrate sustainability risks in their operations and set sustainability objectives. One concern is that it currently focuses mostly on environmental factors, largely ignoring the social impact upon which pioneering European impact investors such as Phitrust and Oltre Venture built their investment theses. As the regulation evolves, it will be important to ensure that it enables the financing of impactful initiatives, not just those with the resources to comply.

Impact investors genuinely interested in creating a positive impact on society need to find the right balance between complying with standards to appear credible and supporting truly innovative companies that are addressing societal problems in a profitable and scalable way. One challenge is the inherent rigidity of impact funds. Investors buy into investment strategies that ideally are valid over the fund's lifetime (around 10 years), but it's difficult to incorporate learnings into such rigid structures, and few impact investors set overarching impact objectives for funds according to a theory of change.

As a result, most impact reports are still aggregations of output numbers (such as "jobs created") or lists of which SDGs the investment strategies address, rather than clear reflections of how the investees generated positive or negative change for the target population. Organizations seldom discuss methodology in their assessments, with the regular excuse that they use a proprietary approach. As impact investors increasingly specialize in issue areas like sustainable agriculture, there's more opportunity to incorporate learnings from previous funds and develop synergies between investee companies. Specialized funds can also better target specific impact objectives and measure success against them.

BUILDING TRANSPARENCY AND A CULTURE OF LEARNING

Grant makers are well-poised for experimentation, risk-taking, and learning to promote social innovation, but they often struggle to match their well-defined missions with solid data and a common understanding of what impact means throughout the organization. Many organizations that work closely with grantees intuitively believe they are changing the lives of their beneficiaries. Rather than burden grantees with lengthy reporting requirements, they may rely on stories about how beneficiaries' lives have changed for feedback. And while grant makers tend to be proud of the passion and trust that permeate their culture, board members from corporate backgrounds may feel that the focus on goodwill and collegiality could generate groupthink and get in the way of transparency.

Boards tend to ask more questions about how the organization manages financial assets than about the impact it is generating for core stakeholders, and there is an inherent tension between upward accountability (how management teams report to trustees and governors) and generating data that informs decisions about impact. If boards encourage management teams to focus on reporting how well they have spent financial assets, they're less likely to share what didn't work well. Ultimately, this inhibits organizations' ability to learn from mistakes, and without risk, there's no social innovation.

Grant makers are increasingly talking about learning as a part of monitoring and evaluation, but not necessarily incorporating it as a concrete activity with associated responsibilities. More of them need to invest in learning units that create a shared language around impact, develop scorecards and tools to track what works and what doesn't work, and train staff to identify the kind of data they need to understand a program's positive and negative impacts on beneficiaries. Good examples of organizations creating physical and temporal spaces for staff to do this kind of work include the King Baudouin Foundation in Belgium, which recently set up an internal learning group to facilitate knowledge sharing, and Laudes Foundation, which has an effectiveness and learning committee as part of its governance structure.

The importance of focusing impact work on accountability to beneficiaries is clear. But to do this effectively, grant makers need to create governance systems that do more to support learning.

FUNDING INNOVATION INTO THE FUTURE

In the coming years, it will be important for impact investors to evolve their impact measurement processes. New standards are beginning to take shape that incorporate social and environmental factors into investment accounts so that they can measure them and subsequently make better-informed

decisions. It's encouraging to see investors rapidly adopting these emerging standards and complying with new regulations, thanks in part to increasingly digitalized measurement tools that facilitate implementation. But the field must work to ensure that the standards and associated tools that prevail aren't necessarily the easiest to implement or the ones with the greatest financial backing. They must be the most equitable and just. They must ensure that investors are accountable to beneficiaries and take into consideration both positive and negative impacts.

Grant makers, meanwhile, need to take impact data more seriously. A lack of solid impact data often means spending vast amounts of resources on projects that have no real impact. To ensure that beneficiaries are at the center of measurement, accounting, and decision-making, more grant makers should consult them as experts and include them in committees and advisory boards. They should also make sure that impact data is at least as prominent as financial data in dashboards and scorecards that inform decision-making. These practices should inform impact standards and regulations for both impact investors and the wider financial markets.

Ultimately, all funders need to use data more effectively to understand which initiatives are generating impact and which aren't. Impact investors speak the language of metrics and can take the lead in translating them into understandable and actionable terms. Grant makers can meanwhile lead the way in making space for experimentation and learning. And through greater collaboration, both groups can help develop standards and accounting systems with true accountability toward and inclusion of those they aim to help.

Lisa Hehenberger is an associate professor at Esade Business School in Barcelona. She established the Esade Center for Social Impact to conduct research with reach, relevance, and rigor for and about social impact.

[1] Lisa Hehenberger, Anna-Marie Harling, and Peter Scholten, "A Practical Guide to Measuring and Managing Impact," European Venture Philanthropy Association, June 2015.
[2] "New European and International Sustainability and Impact Standards," ESIMPACT, 2022.

FIXING THE *S* IN ESG

How to move from net zero to net impact.

BY JASON SAUL

Is the planet really *more important* than the people? According to CNBC, most money managers who use ESG (environmental, social, governance) factors in their investment analysis have focused on the E, or climate change, as their leading criteria for their decisions. But what about the S, or social dimension of corporate impact? As one fund manager put it to me in a recent conversation: "Planet isn't necessarily more important than people, it's just easier to measure. Investors like measuring things that they can put into their models, and carbon is easy to quantify."

No doubt, quantifying social impact is a challenge. A 2021 Global ESG Survey by BNP Paribas revealed that 51 percent of investors surveyed (covering 356 institutions) found the S to be the most difficult to analyze and embed in investment strategies. The report concluded: "Data is more difficult to come by and there is an acute lack of standardization around social metrics. … Investors have been willing to accept data that does little to actually assess the social performance of the companies in which they invest." For most investors, S is merely a check-the-box exercise. So, what can be done to improve S data?

THE MEANING OF S

First, we need to better understand how the field currently defines S. Commentators and investors have described S in many different ways: as social issues, labor standards, human rights, social dialogue, pay equity, workplace diversity, access to health care, racial justice, customer or product quality issues, data security, industrial relations, or supply-chain issues. S&P, one of the leading ESG ratings agencies, describes the S in terms of social factors that pose a risk to a company's financial performance.

In a blog post titled "What is the 'S' in ESG?," S&P outlines three types of S issues:

- **How can a company's workforce requirements and composition present problems for the organization in the future?** Labor strikes or consumer protests can directly affect a company's profitability by creating a scarcity of skilled employees or controversy that is damaging to a corporation's reputation.
- **What risks come with the safety implications of a product or the politics of a company's supply chain?** Corporations that ensure their products and services do not pose safety risks, and/or minimize the

exposure to geopolitical conflicts in their supply chains, tend to face less volatility in their businesses.

- **What future demographic or consumer changes could shrink the market for a company's products or services?** Complex social dynamics, from surges in online public opinion to physical strikes and company boycotts by different groups, affect long-term shifts in consumer preferences. Decision-makers can consider these as important indicators of the company's potential.

It's all a bit of a hodge-podge. The purported through-line, as S&P puts it, is "relations between a company and people or institutions outside of it." That's a pretty ambiguous definition that can cover a lot of things. One could argue this lack of precision in clearly defining S is a major reason why it's so poorly measured.

But there's a deeper existential issue going on here. Rating agencies like Moody's and S&P view ESG almost exclusively through the lens of materiality (i.e., information that is impactful to a company's financial performance). That makes sense because the bread and butter of those agencies is rating corporate and municipal debt, and the primary concern of any investor with respect to debt is, of course, repayment. Risk analysis focuses on the likelihood of repayment. The problem is, most of the interest in ESG is not from lenders evaluating credit risk; it's from investors evaluating equity risk. And equity investors seek to maximize their returns, not just mitigate their risks. Indeed, simply de-risking ESG exposure is unlikely to help investors make affirmative bets on which companies will outperform the market. As State Street Global Advisors noted, "ESG information tends to be the most effective at identifying poor ESG firms that are more likely to underperform as opposed to predicting future outperformers."

The very nature of social impact isn't just about risk; it's also about prosocial behavior. In other words, a company's actions, policies, and investments can and should positively impact people's lives. Of course, there are social impacts like human-rights violations, labor relations, and supply-chain risks that can negatively impact a company's license to operate and financial stability, and those are important. But there are also many social impacts that can positively affect a company's financial performance through competitive advantage, business growth, market relevance, brand purpose, and securing license to operate. Positive social impacts are not accounted for in today's ESG data. Yet, as Larry Fink pointed out in his 2019 letter to CEOs, profits and social impact are "inextricably linked."

Michael Porter, George Serafeim, and Mark Kramer, in a 2019 article titled "Where ESG Fails," argued that "investors who [want to beat the market], as well as those who genuinely care about social issues, have clearly missed the boat by overlooking the significant drivers of economic value arising from the power of social impact that improves shareholder returns."

SOLVING FOR S

To be relevant, the ESG field must modernize the way it measures S factors. To do so, we must overcome several key conceptual challenges: standardization, quantification, and reporting.

Standardization | One of the biggest challenges in measuring social impacts has been the absence of a reliable, quantitative measurement standard. The result is that every company (and NGO) defines, measures, and reports every social impact differently. For investors, this results in unreliable, incomparable, and low-value data that cannot be used in financial models. While there have been a few attempts to create frameworks for reporting social impacts, most have fallen short.

The United Nations' SDGs (sustainable development goals) is among the most prominent of these purported frameworks. However, a 2018 KPMG study titled "How to report on the SDGs: What good looks like and why it matters" found that only 10 percent of companies surveyed had set specific and measurable (SMART) business performance targets related to the global goals, and less than 1 in 10 companies (8 percent) reported a business case for action on the SDGs. Why is this the case? SDGs are primarily designed to track national, population-level statistics such as "mortality rate attributed to unintentional poisoning" or "reduce the global maternal mortality ratio to less than 70 per 100,000 live births." SDGs were not designed to be directly attributed to any discrete social program or intervention. In addition, the SDG goals were intentionally designed to advance the UN's agenda of global development by focusing attention on high-priority topics such as over-fishing, poverty reduction, sustainable tourism, clean water and sanitation, reduced illicit arms flows, etc. While these may be important political goals established by the UN, they are not universally relevant to all companies and all communities.

The ESG field needs an objective standard for reporting social outcomes. Outcomes-based standards are designed to measure the quantum of social change that was realized as a result of a program, strategy, or intervention. An outcomes-based S standard could be used voluntarily by companies and NGOs to self-select which outcomes they want to report against. Investors could also use outcomes data to conduct more robust social impact analysis. For example, investors might analyze whether the impacts generated were in a company's headquarters community or at large? Or whether the impacts are advantageous to recruitment, business growth, competitive advantage, diversity, innovation, market development, or employee health? What "bang for the buck" or ROI did the company generate on the shareholder funds invested? How did this return compare to other companies or to the industry average? Which populations or communities were most impacted? The power of standardized, comparable social impact data gives rise to a whole new level of S analytics that is more incisive, precise, and relevant.

Quantification | Once social impacts are standardized and classified, they must be properly quantified. In the E world, independent bodies like Verra define standards for measuring "units" of environmental impact such as greenhouse gas emissions. Verra refers to these standard units as Verified Carbon Units, or VCUs. Rigorous rules and methodologies are established to ensure consistency and reliability of data across heterogeneous projects. For example, a 1.6 MW Bundled Rice Husk Based Cogeneration Plant in India is measured against the same outcome of VCUs as the Afognak Forest Carbon Offset Project in Alaska.

Social outcomes could be quantified in a similar way. Standards should set thresholds for what constitutes a "unit" of impact for outcomes like hunger, education, and employment. Similar to how carbon credits work, an "impact developer" (i.e., company, NGO, or social enterprise) could report data and have their results verified against the standard. For example, a company might claim that it has helped 1,000 families become "food secure" by providing evidence that each family has achieved the threshold level of criteria for that outcome (i.e., ongoing access to healthy, nutritious food, in a reasonable proximity to their home, on a free or affordable basis).

There's no need to wait for rating agencies to catch up for standard-setters to adapt. Companies have their own independent fiduciary duty to measure and disclose material S information to their shareholders.

Using such a standard, ESG analysts could easily roll-up and aggregate a company's total impact on society. Investors and other stakeholders could actually assess the level of contribution of a business to a critical social issue. Companies could be compared by industry or across industries. Quantification could also be used to price and benchmark social impact. Imagine being able to put a value on a unit of social impact, and eventually trading social impact credits much like carbon? As Scott Kirby, the CEO of United Airlines noted recently in a CNBC interview: "If you put a price on carbon, the public markets will figure it out." It's time we set a price for S and let the public markets figure that out too.

Reporting | In the traditional ESG paradigm, reporting is all about disclosure of "material" risks. But as many researchers have pointed out, there are both negative and positive aspects of materiality. Some activities create material risks that could negatively impact corporate performance and merit disclosure. At the same time, some corporate activities create material benefits that could positively impact corporate performance. In fact, the view that materiality only means material risk is inconsistent with the way mainstream financial markets define the concept. Relying on a long history of existing legal precedent, the SEC defines information as "material" under its Selective Disclosure and Insider Trading Rules if there is "a substantial likelihood that a reasonable shareholder would consider it important" in making an investment decision. There's no suggestion that only risks or negative factors qualify for disclosure. Indeed, many insider trading lawsuits initiated by the SEC are based on materially positive information that contributed to substantial financial gains.

To improve S reporting, the ESG field must expand its view of materiality. The Sustainability Accounting Standards Board (SASB) initially created its "Materiality Map" in 2014 to help investors identify ESG issues that could negatively affect a company's financial performance. Today, the ESG market also needs an "Impact Materiality Map" to help investors identify ESG impacts that positively affect a company's financial performance. An Impact Materiality Map could help investors determine which social impacts are most strategic and beneficial to companies by industry. For example, improving the STEM education pipeline could materially impact innovation and growth in technology firms. For retail grocers, food security and sustainable agriculture could materially influence topline sales. For financial-services companies, financial inclusion can materially expand their customer base and market penetration. For health-care companies, social determinants of health can materially influence their cost structure and patient well-being. And so on. These social impacts are every bit as "material" an influence on corporate performance as risky social issues. Positive social impacts can also serve as risk mitigation for risky social issues. Take Diversity, Equity & Inclusion (DEI) impacts for example. A company's affirmative investment in DEI outcomes (not just box-ticking employee numbers) can have a significant impact on mitigating talent loss and reducing risks to company reputation. These impacts are potentially more material than climate-change reduction to financial-services firms, who already face significant reputational risk in Black communities.

Some social impacts are emerging as universally material to all companies. These can and will change over time, depending on social and political dynamics. Among these "social impact macro factors" are:

1. Public health and its social determinants. If the COVID-19 pandemic taught us anything, it's that major public health crises can affect every business, every industry, and every geography. How companies respond to and address public health needs can be hugely influential over business survival and success. A related issue is health equity, or social determinants of health. The impact of such factors as housing, financial health, and social capital, among others, on chronic illness, employee productivity, and consumer health is directly relevant to all companies.

2. Racial equality. This is more than just a matter of risk and reputation. To compete and grow, companies must focus on inclusivity in their workforce, respond to racism in society at large and make their products and services equitably accessible to all communities. It affects sales, business partnerships, government regulation, employee performance, and competitive positioning.

3. Income inequality and financial inclusion. Of all US households, approximately 44 percent or 50 million people are considered low-income, according to Brookings Institution. That's a pretty massive market segment. Globally, that number is even more significant: 71 percent of the world's population remain low-income or poor, living off $10 or less per day, according to Pew Research Center. To grow and prosper, companies must be able to find ways to include these marginalized populations in the economy and expand the reach of their products and services.

4. Workforce development. Developing a diverse pipeline of talent is critical for every industry and every company. It's not just a positive social impact, it's a key barrier to business growth. Cummins, one of the largest diesel- engine manufacturers, can't service its customers in Africa without trained technicians. Boeing can't build more airplanes without STEM graduates coming out of the public schools. And according to Generation T (an initiative of Lowe's Companies), more than three million trade skills jobs will sit vacant through 2028, which will significantly affect the growth of their business. Growing the nation's workforce, particularly by including underserved populations, has a direct bearing on the growth of almost every business.

Social impacts can also be more or less "material" for different stakeholders: employees, customers, suppliers, or community members. More analysis of social impact materiality will emerge as this data becomes readily available to the investment analyst community.

THE FUTURE OF S

The markets have become transfixed with ESG, and the demand for more and better ESG data will only grow in the years ahead. The success of E data has laid the groundwork for a thriving carbon market—especially the

voluntary carbon market. It has also proven that intangible commodities can be standardized, priced, and traded. This can and will lead to greater impact on the environment than mere advocacy or philanthropic efforts. Indeed, more money is traded through markets every *day* than is spent by all world governments every *year*. To function efficiently, markets must rely on simple, consistent, reliable data. But that data has to signal something. Statistics devoid of meaning have no influence. It's time that the markets value S as much as E and G. The only thing that stands in the way is better data.

There are three practical steps that ESG investors, rating agencies, and companies can do to elevate the importance of S to the markets:

First, and most importantly, companies should start reporting S impact data consistently. Standards have to start from the ground-up. There's no need to wait for rating agencies to catch up or standard-setters to adapt. Irrespective of these players, companies have their own independent fiduciary duty to measure and disclose material S information to their shareholders. Companies should start voluntarily measuring and report their S impacts and get independently verified. This data can then be included in a company's own sustainability reports and 10-Ks. It can also be reported proactively to rating agencies like MSCI, DJSI, Sustainalytics, Moody's, and others. The corporate sector will have a lot more influence over what standards are set if they start producing the underlying data now instead of waiting for the world to agree on it.

Second, ESG investors should start asking for S impact data and making it a requirement. Impact investors like Forthlane Partners in Toronto, Baillie Gifford in Edinburgh, and Planet First Partners in London are already asking for this data. But the process is still manual, the data being requested is inconsistent, and the S analysis isn't as directly linked to corporate performance as it could be. By banding together around a standard for S, data will flow more readily and with less burden on portfolio companies. Over time, with leadership, other investors will join and ask for the same standard S data. The funds that start using this data early will gain a significant edge over competitors. And they will also likely attract new capital faster than run-of-the-mill ESG funds that struggle to answer the fundamental question of so many investors these days: "What impact is my money having?"

Finally, ESG rating agencies, standard-setting bodies, and data providers should align with a specialized S data provider to up-level the value of their data. S impact data is complex; it cannot be simply captured in a one-dimensional box-ticking survey. Reliable, high-quality S data requires specialized taxonomies, questionnaires, and independent verification. This will also create a whole new level of ESG S analysis—that shared value advocates and academic researchers have long argued for. Using this S impact data, rating agencies and others can now begin to evaluate a company's competitive advantage, growth potential, employee resilience, access to new markets, enhanced value chain productivity, and improved operating environment.

As of today, approximately one-fifth (21 percent) of the world's 2,000 largest public companies have committed to meet net-zero targets. Reducing carbon emissions and mitigating the risks of climate change for investors is a major accomplishment. But to achieve true sustainability, we must also improve the quality of life for the people who live on this planet. We can't manage what we can't measure. It's time we raise the bar on social impact measurement, create better S data and give the market something to price into their models. It's time to go from net zero to net impact.

Jason Saul is the executive director of the Center for Impact Sciences at the University of Chicago and the founder and CEO of The Impact Genome Project.

PLOTTING IMPACT BEYOND SIMPLE METRICS

For NGOs, impact comes in different forms and to track the cycles of social change work, we must think across the tangibility and the speed of emergence of change.

BY NATASHA JOSHI

A few months back, we received an internal letter from the founder of an NGO, reflecting on the relationship between funders and founders in the development sector. An anecdote about impact leaped out, in which the community was asked what had changed as a result of the organization's intervention. The founder had expected the community to talk about the NGO's flagship program—targeted at improving livelihoods and income—but in village after village, the community partners explained that "the fear has gone. We are no longer afraid."

How does one measure fear or its *absence*?How does one measure something *priceless*?

The story provoked us to consider what assumptions might be constraining how we think about impact. Compared with domains like manufacturing, in which the relationship between inputs and outputs is relatively direct and causal, ascertaining the impact of social programs is tricky (which makes it all the more important to learn from our partners). After all, societal problems exist within systems, which are not static but inherently reflexive: the very act of intervention changes the system, which, in turn, requires interventions to change. However, development sector measurement and evaluation approaches often don't reflect this reality: most often evaluations are sequenced to come in at the end of a program, as opposed to evolving *with* the program. Implementation to evaluation is usually a straight line, while social change work goes in cycles. And evaluations typically focus on metrics without fully appreciating the second- or third-order effects a change in those metrics might have on adjacent variables.

This is why traditional evaluation (in which expert judges determine the value of an intervention) is giving way—as Emily Gates described at a recent webinar hosted by CECAN—to a new thinking, in which the evaluator is a co-learner who is developing value. As Rohini Nilekani Philanthropies (RNP) has gotten more interested in trying to best understand and articulate the impact (and learning) of all our grants, we started asking our partners to share how *they* see the impact of their work. How did *they* think a philanthropy should judge its own performance?

Our partners work in many different fields—climate action, biodiversity and conservation, gender, civic engagement, justice, media, and youth engagement among others—and across rural, urban, and tribal geographies, with annual operating budgets ranging from $50,000 to $10-15 million and with teams both lean and large. By taking feedback—via an online survey—from this broad range of organization, we tried to get a comprehensive sense of what the sector as a whole thinks and needs.

When we started to analyze the impact data that around 80 organizations shared with us, we realized that in reporting results as "impact," our partners were making a variety of unstated assumptions, as well as treating certain other things as axiomatic. To make sense of it all, we took a step back and broke down what lay underneath. After first reading and discussing all the feedback that came in—and writing a reflection note to get our own thoughts and reactions down—we worked to categorize our grantee processes, outputs, and outcomes by Engagement, Collaborations, Concrete Actions, and Policy Change, after which two broad axes began to emerge, as the framework that follows plots. (See "Grantee Quadrant" below.)

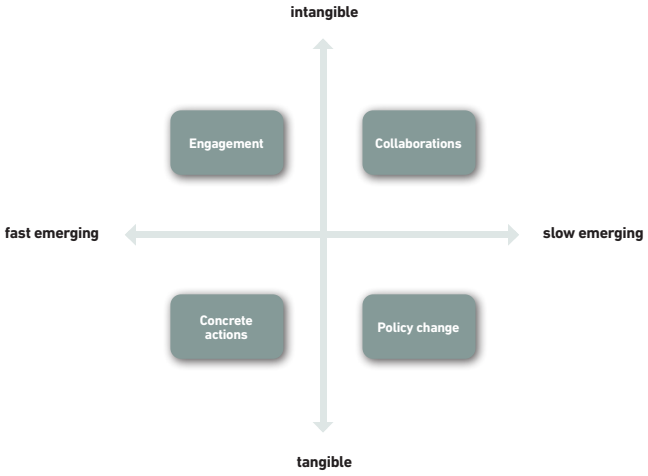
Impact is a spectrum: along the Y-axis, efforts that result in clear policy change are considered tangible outcomes, while the ecosystem collaborations an NGO would forge—in order to get to the policy change—are taken to be "intangible." By the same token, the X-axis plots the results reported by NGOs across a continuum of fast and slow emerging, from quick wins to, well, *slower* wins. Countable actions like vaccinations or children enrolled would be plotted under concrete actions, whereas any progress made toward shifting policies or increasing the network of actors that care about the NGOs' cause come under results that are slow emerging.

What became clear from sorting through the data was that most organizations (if not all) operate in all four quadrants at once, and there is no hierarchy of actions or results. Concrete results are no less important or "strategic" than policy pushes: "countable" results add tangible value to individuals who are being supported on the ground. Field actions also birth insight and innovation that eventually makes it to policy. In the same way, while engagement—under which we have results like "number of report downloads," "number of website/video views," and "frequency with which a network convenes"—can feel ephemeral, consistent engagement is a desirable fast emerging pre-requisite for longer-term deeper collaboration (which is plotted under slow emerging).

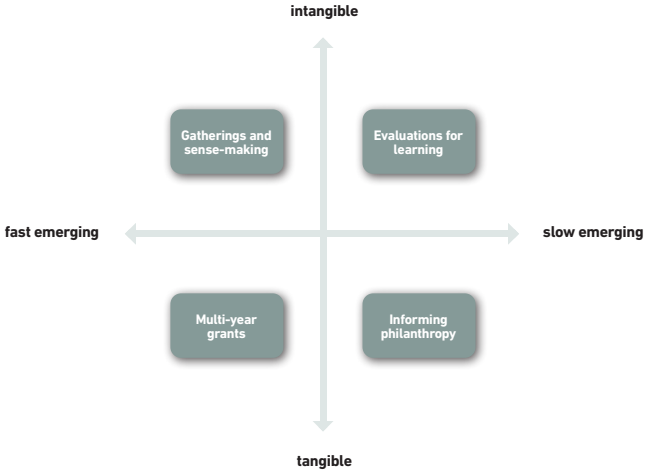
For example, one of our partners in the justice space undertakes concrete actions, providing free legal aid to children in conflict with the law. However, they also collaborate with network organizations that interface with children around other issues; they drive engagement around their work through a series of online and offline outreach events and collaterals (and by engaging volunteers); and, finally, owing to the trust, relationships, and insights built in the field, they are in a position to advocate for reforms to the Juvenile Justice Act (reforms that stand to impact many thousands of children).

Given the two axes exist as continuums, and an organization's work slides up and down and side to side along these axes, it is centrally important to look at how metrics and indicators *speak to one another* across the breadth of the organization's work, instead of focusing on a single isolated metric.

Grantee Quadrant



Donor Quadrant



DEFINING DISTINCT FORMS OF IMPACT

1. Concrete actions: Most nonprofits have a "community" they work with, such that activities done for/with the community can be captured as metrics of progress and impact. However, NGOs often feel the pressure to scale their work—either geographically or tactically (through partnerships or policy change)—out of a desire to see *systemic* change rather than "unit level" change. Many leaders therefore find it difficult to strike the balance between being strategic (through fundraising, building the organization, pushing partnerships, and driving advocacy), and connection to the original cause that brought them into the sector in the first place. Concrete actions are important because working alongside the community and its people is rewarding in a relational way that cannot be substituted. Tangible unit-level actions have been, and will always be, central to social work because that is where the joy and conscience of this work lies.

2. Engagement: Partners typically report engagement metrics as a sign of progress, assuming that increasing interest, awareness, and engagement bodes well for the mission. Be it website analytics, report downloads, likes/hits on digital content, increase in the size of a network, increase in the frequency of interactions between members of a network, and other similar indicators, most organizations see engagement with the larger ecosystem (or with the general public) as desirable. However, depending on the organization's goals, lesser (but repeat) engagement can be preferable than lots of one-time use/engagement (or vice versa!). In the for-profit universe, payment is a good proxy for demand or a person's interest in a product or idea. But in the nonprofit world, where products/services are not measured in terms of payment, should stakeholder buy-in

be instead measured in terms of how much time, effort, or social capital stakeholders expend?

3. Collaborations: Partnerships emerged as the hardest impact metric to define. Organizations reported collaborations and partnerships as positive results, but they frequently defined collaboration very differently. For some it was numeric—more members in a network, more participants on a platform, more collaborators, or the launching of more collectives—while others talked about convergence of efforts by different organizations toward the same goal. When it came to asking why partnerships and collaboration are necessary, the thing we heard the most was that the scale of the problem was too large to be taken on by any one organization. Collaborations bring diverse approaches into the mix, building a critical mass, or collective action. But while this rationale is often taken as an axiom—even while recognizing the difficulty of collectives truly collaborating—networks often end up competing, and partnerships can remain superficial and capricious (versus congruent and stable).

4. Policy change: Some nonprofits focus a lot more on policy change, a hard-won, but highly valued, result. Understandably, policy change affects millions of people and opens up the space for social sector organizations to support the implementation of the new policy, making it a favored goal. At the same time, many organizations see themselves as allies and partners to *other* anchor organizations, who are taking on policy change. Nevertheless, the relationship between field implementation and policy and advocacy can be tightly linked, and organizational focus can shift between one or the other (or include both). Philanthropies too locate themselves in different ways when it comes to influencing policy. Analysis of our own partner data showed us that we tend to favor supporting organizations that aspire to or are playing the role of systems convenors.

HOW MIGHT A PHILANTHROPY GAUGE ITS OWN IMPACT?

Across 80 organizations, certain points came up again and again, and as we sought to reimagine what makes for impactful action, we developed four broad ways forward that linked different forms of impact, which we mapped to the original impact quadrant. (See "Donor Quadrant" on page 17.)

- **Enable CSOs to sustain and grow through multi-year grants:** Most social sector organizations seemed to suggest that the health of civil-society organizations is something philanthropies have a responsibility to foster and build. In action, this could take the form of giving more core, multi-year grants, bringing more donors into a thematic area, funding more diverse and grassroots organizations and leaders. These actions are quantifiable, moreover, and can be engineered relatively quickly.
- **Convene gatherings to build sense-making at the level of the field:** Our partners shared that philanthropic organizations have a vantage that allows them to see intersections between portfolios and domains, and that can help engender thematic or geographic coherence when it comes to tackling systemic problems. Coming together drives engagement with the partner ecosystem as well as between partners, which over time can result in more collaboration and convergence. We plotted this as a fast-emerging, intangible indicator for ourselves.
- **Evaluations for learning:** Formal "evaluations for learning" that results in richer perspectives and shared understandings was seen as the longer arc of a philanthropy's work. Through evaluations for learning, partners, funders, and evaluators stand to see the system as a whole and more clearly. The value of being able to see the system together is hard to quantify (intangible) and also takes time (slow emerging) but is worth investing time and effort in.
- **Drive richer conversations on impact:** The difficulty of sussing out the impact of programs has been written and talked about in many forums. In particular, the (negative) role of donor organizations in pushing a highly metric-focused approach has caused consternation among many NGOs, especially when one metric is valued more than another. As Mona Mourshed puts it, "nonprofits too often receive (well-intended) guidance from stakeholders like funders and board members to disproportionately zero in on a single goal: serving the maximum number of beneficiaries." Given the power dynamics in play, our partners highlighted that RNP and other philanthropies could explore ways in which the discourse on impact could be enriched and expanded, thereby informing the future philanthropy better. This will undoubtedly be a collective journey (slow emerging).

As we search for better ways of measuring progress, Fields of View—a Bangalore-based organization that makes systems thinking actionable through tools—shows what this can look like by distinguishing between "events" and "processes." Events are similar to what we termed "concrete actions": x children enrolled, y acres replanted, z ration kits distributed and so on. Monitoring and Evaluation has matured to the extent that it captures these indicators well. But it fails to account for the processes that bring people into these problems in the first place; if one can see the system better, so as to intervene at the level of process, we stand to change the trajectories of people and problems currently stuck in a loop.

By deploying a systems-conscious, multi-method approach when it comes to capturing and interpreting field dynamics, we stand to increase the resolution of the picture that emerges. This is not to say that doing it is simple, of course. Running evaluations for learning is indeed time-consuming and requires engagement and co-creation. But with emergent technologies, we can start to open up new ways of sensing. Having better sensing tools for impact also allows funders to fund more as this approach reveals many more spaces that require support and funding. Fields of View put it well: "[As] most donors would say, thirty years later we don't want to be funding the same thing."

Natasha Joshi is a development sector professional who has worked with multilateral organizations, foundations, and governments across India, Mexico, and Singapore, and currently leads grants and strategy at Rohini Nilekani Philanthropies. She holds a degree in human development and psychology from Harvard University.

"Nonprofits too often receive (well-intended) guidance from stakeholders like funders and board members to disproportionately zero in on a single goal: serving the maximum number of beneficiaries."

BEYOND RCTs

How J-PAL and the Evidence-to-Policy (E2P) community are integrating innovation and evidence into social policy and practice at scale.

BY IQBAL DHALIWAL, JOHN FLORETTA & SAM FRIEDLANDER

The 2019 Nobel Prize in Economics, awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer for "their experimental approach to alleviating global poverty," is a testament to the innovations they spurred in development economics over the past two decades. But the laureates have repeatedly emphasized that the use of randomized control trials (RCTs) in development economics is part of a broader movement of integrating innovation and evidence into social policy and practice. Equally important to the research was seeding an evidence-to-policy (E2P) community, in parallel, that helped and will help the laureates' research, as the awarding committee put it, "dramatically improve our ability to fight poverty in practice."

The Abdul Latif Jameel Poverty Action Lab (J-PAL), which Abhijit and Esther co-founded in 2003, and of which Michael is a long-time affiliate, now includes almost 500 researchers and more than 400 staff in many countries and is one of the key nodes in a network of E2P organizations and individuals working to advance the use of data and evidence to better understand underlying policy challenges, and to design, pilot, evaluate, and scale innovative solutions for these problems. J-PAL, along with partner organizations such as Innovations for Poverty Action (IPA), have built the infrastructure to fund and implement RCTs, to conduct policy outreach based on insights from this research, and to build the capacity of stakeholders to apply this evidence to policymaking. By leveraging a research infrastructure that includes more than 30 J-PAL and IPA offices around the world, a large and growing network of affiliated researchers have partnered with social innovators in NGOs and governments to rigorously evaluate the impact of promising anti-poverty programs through almost a thousand RCTs in more than 50 countries and in almost all sectors of development, including agriculture, climate, education, firms, gender, governance, and labor.

More than 400 million people have been reached by programs that were found to be effective by researchers in J-PAL's network and were then scaled up by our partner organizations. A stand-alone policy outreach team of J-PAL staff, spread across almost a dozen countries, not only summarizes and synthesizes research evidence into actionable policy lessons, but builds long-term partnerships with local governments and NGOs to scale up effective programs. Thousands of researchers and policy makers worldwide now commission RCTs or use the evidence from them to inform their decisions. This has required dedicated training teams who have created a suite of in-person and, increasingly, online courses to build the capacity of thousands of decision-makers in governments, NGOs, foundations, and other development organizations to conduct RCTs and interpret their results.

Each of these three foundational pillars of J-PAL—field research, policy outreach, and capacity-building—required flexibility, a deep understanding of local contexts, and close collaborations. Above all, a long series of innovations were required not just in the methodology and the econometrics behind RCTs, but also in: data collection in the field (for example, innovative survey design to elicit accurate information and maximize responses),

experimental design (how to minimize spillovers and attrition), transparency (a trial registry, replications, and data publications), scalability (evidence synthesis, replications, a generalizability framework, monitoring systems, and cost collection), and capacity-building (the use of custom trainings and online courses).

Sadly, evidence-informed decision-making is still the exception rather than the rule. Daunting challenges persist in poverty, equity, and increasingly in climate change. In this article, we dive deeper into the key innovations that were behind the success of J-PAL, the learnings from some of these, and new innovations that we believe will be central for us to take on the challenges facing us—while recognizing of course that there are many more dimensions on which dozens of E2P organizations are working and innovating every day.

RESEARCH INNOVATIONS

A unique aspect of J-PAL's role within the E2P community is our focus on methodology. As noted by the Nobel Prize committee, RCTs are a particularly rigorous method of measuring the impact of social programs and determining the causal mechanisms to explain why and how policies do or do not work. But relevant research in the field includes more than just RCTs. Qualitative work is essential to design good randomized evaluations, and descriptive research, administrative data, and continuous feedback from participants are all essential in interpreting and applying insights from

RCTs. Over the past two decades, J-PAL made a series of efforts to help develop a robust research infrastructure by establishing regional offices around the world, supporting hundreds of large-scale field experiments, and improving research transparency and quality.

For example, in response to the challenge of publication bias in the social sciences—the fact that studies showing positive results are more likely to be published—J-PAL helped to establish the norm that new RCTs must be recorded in the American Economic Association's registry for randomized control trials, which J-PAL helped to launch in 2012 and continues to support. During the research design phase, J-PAL encourages researchers to design studies to ensure that the data they collect will be reliable and accurate: reducing spillovers by choosing the most optimal level of randomization (for example, at the entire school level rather than at the classroom level) and minimizing attrition (through stratification). To improve the quality of data during research implementation, J-PAL and IPA helped develop standards for frequent revisiting of data by "back-check" surveyors who reconduct surveys in a sample of households. Additional innovations to improve the quality of data included moving surveys themselves from paper forms to higher-quality digital-data collection with tablets, as well as moving beyond surveys to measure complex outcomes through games, experimental vignettes, and implicit association tests in the field. Post-data collection, J-PAL and IPA have developed robust processes to ensure the confidentiality of participants and the transparent and timely publication of data. J-PAL also provides a replication service to affiliates in which RAs use the original data and rewrite the code used for analysis in order to confirm that the results can be confirmed.

In addition to helping set standards for how randomized evaluations are conducted, J-PAL has also created special funding initiatives to identify pressing policy areas for research and encourage a coherent research agenda. In areas including agricultural technology adoption, crime and violence, education technology and post-primary education, employment, financial inclusion, health-care delivery, governance, jobs, and social inclusion, J-PAL has worked with policy makers and researchers to identify gaps in knowledge and run competitive funds to test promising innovations to help fill these gaps.

J-PAL 2.0

In its post-Nobel phase, one of J-PAL's priorities is to unleash the treasure troves of big digital data in the hands of governments, nonprofits, and private firms. Primary data collection is by far the most time-, money-, and labor-intensive component of the vast majority of experiments that evaluate social policies. Randomized evaluations have been constrained by simple numbers: Some questions are just too big or expensive to answer. Leveraging administrative data has the potential to dramatically expand the types of questions we can ask and the experiments we can run, as well as implement quicker, less expensive, larger, and more reliable RCTs, an invaluable opportunity to scale up evidence-informed policymaking massively without dramatically increasing evaluation budgets.

Although administrative data hasn't always been of the highest quality, recent advances have significantly increased the reliability and accuracy of GPS coordinates, biometrics, and digital methods of collection. But despite good intentions, many implementers—governments, businesses, and big NGOs—aren't currently using the data they already collect on program participants and outcomes to improve anti-poverty programs and policies. This may be because they aren't aware of its potential, don't have the in-house technical capacity necessary to create use and privacy guidelines or analyze the data, or don't have established partnerships with researchers who can collaborate to design innovative programs and run rigorous experiments to determine which are the most impactful.

At J-PAL, we are leveraging this opportunity through a new global research initiative we are calling the "Innovations in Data and Experiments for Action" Initiative (IDEA). IDEA supports implementers to make their administrative data accessible, analyze it to improve decision-making, and partner with researchers in using this data to design innovative programs, evaluate impact through RCTs, and scale up successful ideas. IDEA will also build the capacity of governments and NGOs to conduct these types of activities with their own data in the future.

J-PAL is not alone in acknowledging the significance of administrative data. In the United States, groups such as the Actionable Intelligence for Social Policy Initiative and The Lab @ DC are utilizing administrative data at the state and local government level to better understand the impact of policies on the communities they serve. Internationally, organizations such as Development Gateway and Global Partnership for Sustainable Development Data are working with government agencies to improve the access and use of their administrative data for evidence-based policymaking. Even organizations that historically do not conduct randomized evaluations, such as the International Monetary Fund, recognize the important role of utilizing administrative data to improve statistical quality and support government agencies in their policymaking.

The second big scope for innovations in research for J-PAL and the wider impact evaluation and evidence-informed policymaking communities is building more expansive and ambitious research agendas in particular sectoral areas. One of the most pressing areas for new research is climate change, which threatens to undo decades of progress and whose consequences are already being felt through heat-related mortality, deteriorating food security, and extreme weather events. People living in poverty are the most vulnerable to dramatic changes in temperatures that affect agricultural livelihoods and make it more difficult, expensive, or dangerous to live in locations affected by sea-level change, natural disasters, and extreme temperatures. Without tackling climate change, making progress in solving many other problems will be significantly less impactful—for example, if significant parts of Rio de Janeiro are underwater due to rising ocean levels, if the air quality is too dangerous to go to school in Delhi, if fires disrupt travel and infrastructure in Australia, or if it's simply too hot to go to work in Saudi Arabia.

With one foot in research and another in policy, the E2P community is well-placed to tackle this issue. First, researchers must test technologies

Researchers must test technologies from the lab in the field. Measuring the real-world impact of new technologies is crucial to understanding the political economy factors that will affect their potential to scale.

from the lab in the field. Measuring the real-world impact of new technologies is crucial to understanding their externalities and the political economy factors that will affect their potential to scale. Second, researchers need to test strategies to improve the effectiveness of climate policy and regulation, as some of the countries that will be hardest hit by the consequences of climate change also grapple with limited state capacity to design and enforce regulation. Finally, researchers must test interventions that affect behaviors in order to reduce individual burdens and encourage certain behaviors.

An example of this work in practice comes from Gujarat, India, where making environmental auditors more independent improved the accuracy of pollution audit reports, leading, in turn, industrial plants to pollute less. Based on the results of a randomized evaluation by J-PAL affiliates, the Gujarat Pollution Control Board (GPCB) reformed its environmental auditing system in 2015, issuing new guidelines that require random assignment of environmental auditors. Members of the research team continue to work closely with officials in Gujarat and other Indian states on environmental policy design and evaluation.

Refugee populations and migration is another vital area for innovation in experimental research to inform policymaking. The United Nations estimates that there are now more than 70 million people worldwide who have been forcibly displaced from their homes, the highest figure recorded in history of refugees, asylum seekers, and internally displaced people. Many experience compounding vulnerabilities: over half of refugees globally are children, millions are stateless and have therefore been denied legal protections, and many come from or live in countries with serious safety and security concerns.

It is vital that the E2P community double down on efforts to learn how to address challenges to education, safety, and employment for populations who can often be displaced for decades. At J-PAL, for example, our European Social Inclusion Initiative evaluates programs and policies to foster the social inclusion of migrants and refugees in Europe. The first round of funded projects includes evaluations of programs that, for example, foster social ties to promote the integration of migrants in Sweden, build social cohesion between Turkish and Syrian schoolchildren, and counteract the social exclusion of immigrants in Finland. IPA also recently launched a Humanitarian and Forced Displacement Initiative to improve the lives of those who have been forcibly displaced and the communities that host them through policy-relevant research.

Other areas where innovative research questions are now being asked include gender, the private sector, taxation, and urban infrastructure. Development economics no longer only studies questions of gender through disaggregated data, but goes much further to ask questions such as: how we can most effectively address gender disparities and inequality at scale, whether existing development programs are closing the gender gap in human development, how gender dynamics in families and society affect the impact of these programs, and how we can best measure changes in areas like agency and empowerment. Researchers such as Lori Beaman, Pascaline Dupas, Erica Field, Seema Jayachandran, Rohini Pande, Simone Schaner, and many others are now designing innovative studies to answer these questions.

Although the private sector may seem to fall outside the bounds of poverty research, there is scope for experimental research. Income differences between countries can be explained largely by differences in firms' productivity, for example, such that identifying policies that stimulate productivity growth or enable high-productivity firms to grow can have important consequences for poverty alleviation and social mobility. Bringing rigorous evidence to conversations about how to generate firm growth—and how this growth affects workers, their families, and the broader economy—can help inform and improve these policies. Researchers addressing these questions include David Atkin, Nicholas Bloom, Dean Karlan, and Antoinette Schoar, among many others. Recognizing the importance of emerging research in these areas, J-PAL formally launched new Gender and Firms sectors last year.

New data allows us to ask new questions, but the challenges faced by the world's poor also continue to evolve. Researchers and E2P organizations must be flexible enough to address these new problems and collaboratively find new solutions.

POLICY INNOVATIONS

Randomized evaluations are valuable in precisely estimating the impact of a particular social program in a particular place and time, but causal impact alone is not always sufficient to improve policies and change lives. Programs and policies shouldn't be scaled up simply because they meet a particular bar of effectiveness. Evidence-informed policymaking, particularly when adapting evidence-backed social programs from one context to another, requires a deep understanding of the global evidence, knowledge of the local context and local systems, a window of opportunity, political will for change, funding, and sufficient implementation capacity.

Policy makers face multiple constraints to using evidence to inform their decisions. On the "supply of evidence" side, these include challenges to accessing relevant research, the fact that impact evaluations rarely include detailed program cost data (budget constraints are as important for a policy maker as impact), and the difficulty of determining actionable take-aways when evaluations of similar programs show different results. Furthermore, will a program that was successful in one location achieve similar impacts in an entirely new context? On the side of "demand for evidence" from potential users, we also need to appreciate that policy makers need to make decisions on short timeframes and must factor in a range of political, administrative, and budgetary considerations. Evidence is, at best, just one part of the decision-making process. In working to translate experimental evidence into action, a key lesson we have learned is to work on both the supply and demand sides.

We have also learned to apply judgment about the types of tested policies and programs which should be scaled in different contexts and to build coalitions and long-term partnerships for change. J-PAL's evolution toward basing policy staff in developing countries and working closely with local policy makers to apply insights from research happened gradually. We started by trying to make it easier for policy makers to access relevant experimental evidence on program effectiveness, the "supply problem." We expanded the pool of policy-relevant research by launching new initiatives to spur projects in areas such as governance and urban services, requiring researchers to collect cost data in addition to impact data. To make the findings of randomized evaluations more accessible to a non-academic audience, the policy group began to produce two-page synopses of the relevant policy questions, the contextual factors, the implementation details, and the results of the programs being tested. These are written in a non-technical style and are available in a searchable database on the J-PAL website. To highlight evaluations that addressed particularly relevant questions for policy makers, we created glossy,

expanded summaries called briefcases. To provide a bigger picture on the implications emerging across multiple studies in one thematic area, we developed syntheses called bulletins, which begin to reconcile conflicting results and draw coherent and actionable policy lessons. Many bulletins include cost-effectiveness analyses calculating the ratio of the impact each program achieves to the cost incurred to achieve that impact. Over time, to make these syntheses even more digestible and timely for policy makers, J-PAL developed shorter, two- or three-page policy-insight products to highlight the emerging consensus on policy lessons from J-PAL academic co-chairs. To help policy makers appropriately apply these lessons to their contexts, we developed a framework on how evidence and interventions can generalize from one place to another.

However, we realized that even all of these collective efforts were insufficient to fully leverage the potential treasure troves of research—we needed to equally address the "demand" side of evidence-informed policymaking. To complement the policy teams developing these products at J-PAL's global office at MIT, we hired policy staff based at J-PAL's regional offices located within local universities around the world, whose job is to work directly with national, regional, and local policy makers to help them understand and apply experimental research.

Perhaps J-PAL's most important learning centers on how to effectively create partnerships to scale tested innovations with governments. To many people, governments—particularly developing governments—seem ineffective, corrupt, or slow to innovate. However, governments are the pre-eminent players in global poverty and development, with unrivaled responsibility for social programs and incomparable reach and scale, and big gains in welfare can be catalyzed through smart partnerships. Over the last 20 years, 400 million people around the world have been reached by programs that were scaled up after they were shown to be effective through evaluations by J-PAL researchers. The vast majority of those scale-ups were led by governments.

We have learned three key lessons in how to build partnerships with governments to scale evidence-backed innovations:

1.

"Globally informed, locally grounded" policy staff can identify and leverage policy windows.

J-PAL policy staff based in regional and country offices or embedded in government departments have been critical in identifying windows of opportunity for evidence to inform decisions, as well as finding "evidence champions" within governments who can help make that happen. In Zambia, for example, frequent visits by the J-PAL Africa policy staff to understand the priorities of the Ministry of Education led to an opportunity to contextualize and scale up Teaching at the Right Level (TaRL), an approach pioneered by the Indian NGO Pratham and developed iteratively through multiple J-PAL evaluations led by Banerjee and Duflo. Through these interactions, J-PAL staff learned that Zambia was wrestling with low learning levels in primary schools and was looking for ideas to help students catch up after falling behind.

2.

E2P organizations can support governments in understanding and adapting tested ideas to their context.

We need to make it as easy as possible for governments to understand the mechanisms of what made policies work in other contexts and whether and how those programs could apply to their setting. In the Zambia case, we helped the Ministry of Education understand the different evaluated models of TaRL and the essential local conditions and implementation details to making the program work. We traveled with them on a "learning journey" to India to see how Pratham and Indian governments implemented the program. When they were convinced, we invited Pratham to Zambia to help them develop their own materials and models.

3.

Coalitions are critical to supporting scale and sustainability.

Even after contextualizing and piloting, governments have requested and benefited from support in integrating the model into their systems. This includes developing monitoring systems (including improving the government's own administrative data), ensuring fidelity to core components of the original model, fundraising, and scale-up planning. E2P organizations such as J-PAL have helped develop coalitions of NGOs, researchers, and funders to support governments during this stage—achieving together what no organization could accomplish independently. In Zambia, the ministry leads implementation with J-PAL, Pratham, and the Flemish NGO VVOB providing implementation and monitoring and assistance from a group of funders led by USAID. The model is now improving literacy and numeracy for children in 1,800 schools across the country. With support from Co-Impact, this work is being extended to other African countries with the aim of reaching three million students over the next five years.

How will J-PAL apply this learning about catalyzing evidence for policy action from the last decade? First, we are doubling down on scaling through building multi-stakeholder partnerships with governments through a competitive fund. J-PAL's Innovation in Government Initiative (IGI) works with governments to adapt, pilot, and scale evidence-informed innovations that have the potential to improve the lives of millions of people living in poverty in low- and middle-income countries. We are running IGI as a competitive fund to which J-PAL regional offices and partners, in collaboration with governments and researchers, can apply for support to help seed evidence-backed innovations for scale. In particular, the initiative is promoting the use of technology and data-enabled systems to reduce the costs of program delivery and monitoring, implementation

We need to make it as easy as possible for governments to understand the mechanisms of what made policies work in other contexts and whether and how those programs could apply to their setting.

science to adapt and pressure-test different models before scaling, and the systematic collection of cost data.

A second big policy theme of "J-PAL 2.0" will be to better leverage policy windows, the openings when partners are more open to informing decisions with evidence. J-PAL plans to prioritize more nimble, flexible, and fast support to partners and strategic organizations by sharing evidence, tested policies and programs that may be most effective in different contexts, and the critical mechanisms and details to making them work, as well as curating discussions between the researchers and implementers. We envision these opportunities may arise at specific times when organizations are revisiting their strategies or entering new multi-year planning cycles. For example, Ben Olken, J-PAL Co-Director, and J-PAL policy staff shared global evidence and frameworks for how to leverage philanthropy to improve governance with the William and Flora Hewlett Foundation as an input to its strategic planning. On April 16, 2020, Abhijit will present to 50 Finance Ministers during the World Bank's spring meetings to share tested ideas for improving human capital through education, health, and social protection.

Finally, a third theme for the next decade of policy outreach at J-PAL will be to translate these most promising collaborations with policy makers into durable, long-term partnerships. Long-term partnerships—with other E2P organizations such as IPA, governments such as Peru's Ministry of Education and the Indian state government of Tamil Nadu, NGOs such as Pratham, and donors such as the United Kingdom Department for International Development or the Bill & Melinda Gates Foundation—have been essential in contributing to a movement of evidence-informed policymaking. Many of the scale-up examples cited above were made possible by the trust and mutual understanding that arises through long-term partnerships.

CAPACITY-BUILDING AND DIVERSITY INNOVATION

To empower policy makers and funders with the tools to incorporate evidence into their work, J-PAL and many E2P organizations such as the World Bank and the Center for Effective Global Action (CEGA) have prioritized capacity-building efforts. J-PAL helps to develop the capacity of researchers who produce evidence and the policy makers and donors who use it. We have done this through a range of activities, from in-person training programs to university-level open online courses.

Over the past two decades, in addition to internal training for research associates, J-PAL created a suite of training courses, including our flagship training event: Evaluating Social Programs, a five-day executive education program that provides hands-on training to policy makers on randomized evaluations and becoming better commissioners and users of evidence. These open-enrollment courses are complemented by a range of custom in-person courses for partners such as the European Commission, the Indian Administrative and Economic Services, and Unicef. Since 2003, we have reached more than 8,500 participants.

WHAT'S NEXT?

As we look to the future, three interrelated areas will more effectively increase the capacity of our partners and spur more evidence-informed decision-making. First, we will leverage the potential of online platforms to reach more people; second, we will integrate these courses into government civil service trainings; and third, we will help change the field by working with a greater diversity of researchers from different backgrounds.

In its early days, J-PAL relied heavily on in-person courses to train staff and partners. However, online platforms have the potential to dramatically increase access to this knowledge, such as the online MicroMasters program in Data, Economics, and Development Policy that we offer in partnership with MIT Economics. Though not the first standalone online courses on development, this program was the first of its kind to offer a fully online program grouping together a series of courses on policy and poverty to enable participants to earn an official MicroMasters credential: through five online courses and in-person exams, the program equips learners with the practical skills and theoretical knowledge needed to tackle pressing challenges in social policy. To increase access, the course is affordable: the cost varies depending on the annual household income of the learner, from $100 to $1,000 per course, and can be audited for free if the learner does not wish to sit for the certificate exam. The courses are offered several times a year, and learners can take anywhere from one to five courses at a time. The pricing structure, timing and pacing flexibility, quality, and online nature of the course allow for high participation worldwide: To date, more than 10,000 learners from 180 countries have signed up for a course, and more than 1,500 have earned an official certificate of course completion.

The second future priority for J-PAL is the lifelong and tailored learning of civil servants and government officials. Unfortunately, civil servants in many countries receive either very little training (concentrated around when they are hired) or receive training at very discrete steps in their career (for example, at five-year increments). Rather than conducting a series of one-off custom trainings with government departments, we will leverage online-learning modules we are developing to prioritize long-term, institutional collaborations with government civil-service programs. For example, under its partnership with the J-PAL South Asia regional office, the Government of India's Department of Personnel and Training is now offering reimbursements for eligible staff to complete any of the MicroMasters courses; 67 civil servants signed up for classes in the pilot phase, and 38 of these have completed a course and passed the final in-person exam so far. Our expectation is that thousands more will join over the next few years. We are now working to expand the number of course offerings to dozens of options across numerous government departments with support from the Bill & Melinda Gates Foundation.

A third important priority related to capacity-building is working with and strengthening local researchers in developing countries as part of a broader effort to diversify the worldwide researcher pipeline. Economics researchers often come from developed countries and elite institutions. This is true in the field generally as well as within the J-PAL affiliate network specifically. In addition to issues of equity and opportunity, this does a disservice to our wider goals of evidence-informed policy and poverty alleviation, as local researchers offer unique insights and perspectives on the challenges and potential solutions to poverty-related issues grounded in their knowledge of local context.

At J-PAL, the highly accessible MicroMasters courses are an important component of efforts to diversity research networks and invest in developing local capacity. Another example comes from the J-PAL Digital Identification and Finance Initiative, hosted by J-PAL Africa. The initiative's director, Tanvneet Suri, and initiative staff have established pilot funding for African researchers, are integrating local researchers into the initiative as lead investigators and co-authors, and are providing targeted mentorship from senior affiliated researchers in our network. Other initiatives to better integrate researchers from developing countries

include PhD fellowship programs, academic conference scholarships, and "matchmaking" efforts through other research initiatives to match local researchers with evaluation opportunities.

Gender diversity within the field of economics is also a priority, as women remain underrepresented in many aspects of the field. J-PAL recently established a Gender Working Group to examine internal issues concerning gender, recently held implicit bias training for staff, and is actively discussing ways to include more women in our network of research affiliates.

CONCLUSION

As the scale of challenges facing those living in poverty evolves to encompass new threats from rising sea levels and temperatures, new magnitudes of migration and displacement, and new realities in dense urban centers, the way that policy makers and researchers ask and answer questions must shift as well. The potential role of evidence to improve social policy has never been more important. To help face these challenges, J-PAL and the broader E2P community will continue to push innovations in research, policy influence, and capacity-building. The Nobel Prize's recognition of the work of evidence-informed decision-making is a major boon to the broader movement and will help us to more ambitiously push for change and to increase the role of research in reducing poverty.

Iqbal Dhaliwal is the Global Executive Director of J-PAL. Based at MIT's economics department, he works with the Board to develop the organization's strategic vision, and with the leadership of the regional offices to coordinate research, policy outreach, capacity-building, and operations worldwide. He is also the co-Director with Esther Duflo of J-PAL's South Asia office, and with Abhijit Banerjee of the Innovation in Government Initiative.
John Floretta is Director of Policy and Communications for J-PAL. He works with policy makers, J-PAL affiliated researchers, and J-PAL global and regional staff to disseminate lessons from randomized evaluations and promote evidence informed decision-making and scale-ups of successful social programs.
Sam Friedlander is the Senior Policy and Communications Associate to the Executive Director of J-PAL.

TEN REASONS NOT TO MEASURE IMPACT— AND WHAT TO DO INSTEAD

Impact evaluations are an important tool for learning about effective solutions to social problems, but they are good investment only in the right circumstances.

BY MARY KAY GUGERTY & DEAN KARLAN

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.

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 can

help create a better program and stronger organization. These data often get lost or overshadowed in the pursuit of impact evaluations.

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?

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.

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.

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.

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

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

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.

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.

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.

Mary Kay Gugerty is Nancy Bell Evans Professor of Nonprofit Management at the University of Washington's Evans School of Public Policy and Governance.
Dean Karlan is chief economist at USAID.

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.

PUTTING EVIDENCE TO USE

Research does no good if its insights are irrelevant or not applied. Ensuring that evidence influences policy requires developing the right ecosystem and levers for accountability.

BY HEIDI MCANNALLY-LINZ, BETHANY PARK & RADHA RAJKOTIA

Research papers on innovative social programs may be grounded in painstaking and pricey evidence collection, but what good are they if they go unused, not to mention unread? According to a 2014 study by the World Bank, nearly a third of the reports available as PDFs on their website had never been downloaded even once.

International research organizations have made some progress in ensuring that evidence generation goes beyond mere publication and instead reaches and informs decision makers. Over the past 18 years, our organizational strategy at Innovations for Poverty Action (IPA) has evolved from a narrow focus on evidence generation, to a linear understanding from evidence generation to dissemination, ultimately to a grounded and iterative approach of cocreation between researchers and end users of research.

At its core, our strategy identifies the right opportunities for evidence to influence real change, partners with end users to answer their questions, uses a research tool kit that goes beyond impact evaluations, empowers local researchers and decision makers, and invests in localized data capacity to ensure that further learning can be sustained. We are no longer solely an evidence-generating organization, but rather an evidence mediator: We work with a variety of implementing partners to collect data and evidence and put it into proper use.

More specifically, we at IPA apply an integrated framework of the what, who, and how of evidence use. In this way, we ensure that funding for evidence-to-policy work is less about production of papers and more about building meaningful, multidimensional partnerships that ground critical decisions in evidence of effectiveness.

By better understanding how evidence actually gets used, funders—from small foundations to larger ones, to government funders and multilateral development banks—must change the way they invest if they want to realize the potential of evidence. And evidence mediators, like IPA, need to put the quality and depth of our partnerships on the same level as the quality of the evidence that we generate if we want to ensure that evidence is actually used.

WHAT, WHO, AND HOW

Through our work in more than 20 countries with partners at varying levels of experience with data, IPA has learned that encouraging evidence use depends on context, especially on what we call *evidence readiness*: the preexisting and ongoing experiences with evidence, data, and application of research to practice in each country, sector, or institution in which we seek to encourage evidence use.

Our evidence readiness framework ranges from contexts of unreliable data (working with a partner who has almost no access to reliable data) to contexts of rich data and evidence use (partners whose creation and use of data and evidence are regular and ongoing). Unsurprisingly, the vast majority of interested partners fall in a middle range. Partners who already have worked on some concrete examples of applying evidence to program and policy design have the further opportunity to build their own evidence generation and scaling of evidence-informed programs, thereby reducing the role of evidence mediators.

The framework teaches users not to treat each context uniformly and to see opportunities to pursue evidence use in a wide range of contexts. Because of contextual variation, the pathway to evidence uptake is not linear or uniform, but our experience suggests that finding high-impact opportunities (*what*), building the ecosystem to support evidence use (*who*), and leveraging targeted tools (*how*) can help focus efforts on impact.

The what of evidence use: Finding high-impact policy opportunities | Not every opportunity in international development research has strong potential for evidence use. IPA prioritizes opportunities that meet four criteria: an existing body of research to build on, an opportunity to influence important decisions, existing relationships, and existing funding for implementation.

For example, IPA has partnered with the Rwanda Education Board (REB) since 2014. Our prioritization framework applied particularly well when the REB—together with IPA and other partners—took the opportunity of centralizing teacher recruitment and rewriting its human capital strategy to incorporate both evidence and data. This opportunity met all the criteria. A strong body of cocreated evidence around performance contracts for teachers was already in place. In addition, IPA already had

strong working relationships with critical people in Rwanda's education ecosystem. What's more, important decisions about policy were about to be made: Rwandan officials were preparing to rewrite teacher recruitment and deployment strategy. Finally, the evidence-based program would be cost-neutral in the medium term for the government, making funding for implementation a surmountable issue.

The who of evidence use: Equipping an entire ecosystem with evidence | Building a culture of sustained evidence use requires a broader approach than finding one particular evidence champion. It means engaging the whole ecosystem of relevant actors and emphasizing each one's incentives for reaching their own impact goals. This ecosystem includes technical staff across departments and organizations, more senior-level ministry officials and political leadership, and multi- or bilateral funders and their government counterparts.

We have also found that partnering with researchers, policy makers, and practitioners from low-to-middle-income countries, who have the capabilities and insight necessary to generate and apply the most relevant evidence in their context, can accelerate the process. This strategy is typically more fruitful—and equitable—for evidence brokers than privileging their own perspectives or relying on "expertise" that is not grounded in the local context.

The how of evidence use: Clearing a pathway for evidence use, then equipping partners to follow through | Even when the correct opportunities for evidence use are identified and the right coalitions are built, failures to use evidence can sometimes occur when the relevant parties commit to evidence collection but fail to do the complementary work that would actually lead to evidence use. For example, researchers can focus on the causal mechanisms at work in an intervention but fail to collect data on crucial programmatic details about delivery and implementation. Or they can draw up a map of relevant stakeholders but fail to engage them or formulate action plans. Just as it is a mistake to think a single evidence champion can bring about transformational evidence use, it's erroneous to think that partial investments in the how will bring about impact.

For example, IPA and a large group of researchers from Stanford University and Yale University ran a massive randomized controlled trial (RCT) last year on improving mask-wearing in Bangladesh to prevent COVID-19. The model we tested more than tripled mask-wearing, and that effect persisted beyond the intervention. Since this approach had the power to save thousands of lives at very low cost in the middle of a case surge in South Asia, we shifted from research to large-scale implementation very quickly. The first scale-ups—to 4 million people in India by the Self-Employed Women's Association (SEWA) and to 81 million people in Bangladesh by BRAC—needed an urgent, easily deployable monitoring tool to understand if the program could work in different contexts and at scale, and to inform management along the way. If we had not invested in this monitoring, we may not have been able to persuade more partners to take this on—and we certainly would miss critical gaps in implementation at such a large scale.

DOING BETTER

Generating research—whether RCTs, data, or any other kind—is only half the battle and serves no purpose if the evidence collected isn't used. The translation to evidence use—identifying the right opportunities, equipping the ecosystem, and using all the right tools—is currently both unstructured and underresourced. Funders and recipients need to commit to a full evidence-to-policy cycle, in which they have a plan for evidence use and are held accountable to the outcome. To achieve this goal, investments need to be partnership-focused, flexible, long-term, cost-effective, and based on data-driven learning about what works to spur evidence use.

Local partnership-focused | The organizations that create evidence are not always those who use it, and this mismatch can create tricky funding scenarios. But granting agencies can use their leverage to secure partnerships between evidence-creating and evidence-using entities and hold them accountable to completing an evidence-use cycle. They can also ensure that the funding for implementing an evidence-informed program doesn't run out at the very moment that evidence supporting its use emerges—a Kafkaesque outcome that we have experienced all too often. Locally based evidence mediators can broker these partnerships, advocate for effective use of funding, and ensure that knowledge is not lost in the transition between its generation and use.

Flexible and long-term | Funding for evidence use needs to be outcome-focused. It should incorporate flexibility around short- and medium-term outputs, since the pathways to policy change are varied and intertwined, and it also needs to last over the long term to secure the impact sought. Windows for influence over policy can open and close quickly, often through events outside the control of funding recipients. Funding milestones that are too rigid pose barriers to making the most of data and evidence collection in shaping outcomes.

Evidence-informed and cost-effective | Funders and recipients should commit to learning what works for achieving evidence use and pursue evidence-informed, cost-effective evidence-to-policy models. Agencies that support evidence use in global development, such as USAID Development Innovation Ventures, Global Innovation Fund, and the new Fund for Innovation in Development, apply a return-on-investment perspective when thinking about their grants. Nobel laureate economist Michael Kremer, the scientific director of USAID Development Innovation Ventures, and his colleagues have done valuable research into which investments in development innovations have paid off.

We must build on such work by assessing the relative efficacy of various evidence-to-policy strategies. As far as we know, this learning and evaluating is not being done in a systematic way. As an evidence-informed development community, we can do better.

Heidi McAnnally-Linz is the global lead for policy and partnerships for BRAC International's Ultra-Poor Graduation Initiative.
Bethany Park is senior director of policy at Innovations for Poverty Action.
Radha Rajkotia is CEO of Building Markets.

IN SEARCH OF DURABLE CHANGE

Measuring how long impact lasts can be difficult, but nonprofits and donors should make the effort.

BY MONA MOURSHED

As I argued in my 2022 article, "Beyond 'X Number Served,'" nonprofits and donors should expand our thinking beyond the number of beneficiaries we reach. We should hold ourselves accountable for two other key aspects of impact that can be hidden behind numbers or get lost in statistics in yearly reports or grant applications: How deeply and well are people served? And how long does the impact actually *last*?

While I still believe real impact requires delivering breadth, depth, and durability simultaneously, there are particular challenges to measuring impact across time that make it worth thinking harder about. Time and again, I've heard people say that durability is "too hard to measure" or insist that "we are not responsible for impact years beyond initial program delivery" and "no one is asking us to measure it!" And yet, durability is where the rubber meets the road. In health care, expanded treatment doesn't mean much if it doesn't lead to measurable and sustained increases in health outcomes. If we address hunger by providing meals, what we really want is to progress toward stable long-term food security for disadvantaged communities. And so on.

The time has come for nonprofit leaders and donors to launch deep and sustained efforts to tackle durability: the least discussed, least measured, but arguably most important of these three key metrics.

WHERE IS THE DATA?

Let's zero in on the field I know best, employment. At Generation, the global employment nonprofit network that I lead, our goal is to train and place adults of all ages in new careers. But if our employed graduates fail to earn a living wage, or fall out of work a year later, what have we really achieved?

As a workforce field, we have collected shockingly little evidence that the hundreds of billions of dollars spent annually by governments, foundations, corporations, and individual learners on training and reskilling actually results in lasting improvements in income and well-being. What analysis we *have* gathered shows at best a low to mixed return. A 2017 analysis of 12 technical and vocational education and training (TVET) programs across 8 countries examined employment impact 12 to 18 months post-program and found that these efforts, on average, increased employment by only *two* percentage points. Similarly, Mathematica conducted a 2023 review of 17 impact evaluations of TVET programs across low- and middle-income countries and concluded that only 4 demonstrated a statistically significant impact on employment beyond 12 months. What about lifting incomes? J-PAL's 2023 review of 28 randomized evaluations of TVET programs, which also agrees that most programs increase employment only modestly, found that just half increased graduates' earnings *at some point in time*. But a big part of the problem is that we just don't know how durable our intervention has been; as the study concluded, "To date, there is not a very clear understanding of what influences whether an intervention works in the short run in comparison to the long run."

DATA FORWARD

Generating reliable data on durability requires collecting the ongoing results of individual outcomes, long after a specific intervention has ended. This is hard: it means staying in close touch with program graduates to keep pace with inevitable changes in phone numbers and emails and maintaining deep enough bonds that people will be motivated to report back on their status year after year. But doing this requires more than determination. It requires creativity. At Generation, for example, while data completion rates for our alumni surveys start out at 90 to 100 percent within the first year following program completion, they fall to around 60 percent at one to two years post-program, and then settle to around 30 percent after two to five years. Maintaining even that comparably high level of long-term response rate across our now 100,000-plus global alumni is hugely valuable, but it's not good enough. For example, there is a problem of positivity bias: since it's likely that employed alumni will be more inclined to respond than those who are unemployed, we need much higher response rates to reliably speak to the trajectories of diverse learner profiles across the geographies we serve.

To improve data completion at Generation, we are now moving toward multi-channel follow up, which relies on a combination of emails, online surveys, SMS/WhatsApp messages, direct one-to-one follow ups via text or phone call, and in-person meetings at alumni events. We are also exploring using interactive voice response for short surveys about job retention or wages.

This kind of extensive data collection does not have to be expensive. It currently costs Generation 1 percent of the total cost per learner to measure our durability outcomes. However, few philanthropic or government donors incentivize durability data measurement, much less pay for it. Workforce funders, particularly governments, are emblematic in the persistent focus on expanding the number of people served, which means that funding rarely goes beyond covering program-delivery costs, with little to no requirement or support for reporting on program outcomes beyond the grant period.

And yet, wouldn't investing in better long-term impact measurement also sharpen our ability to make operational improvements? It certainly did at Generation, when we began tracking an "impact ratio," or the extent to which Generation is filling a percent of annual vacancies for a target profession in a target city (for example, junior full-stack developers in Guadalajara, Mexico). In 18 locations across 8 countries, we now hold more

than 5 percent of entry-level jobs, which is a significant share of hiring, up from 9 locations in 4 countries one year ago. Our new approach enabled us to identify which professions had the greatest potential for growth, and to build an employer ecosystem to achieve it. What got measured indeed got managed.

More is required than data, however. To tighten the links between better measurement of durability and better management by service providers, we need common datasets. In the employment and training field, for example, living wage attainment should be the gold standard, the most objective measure that a provider can use to assess the economic mobility of graduates over time. But except for the United States, the United Kingdom, Canada, and Australia, robust living wage benchmarks are not available publicly (or regularly) for most countries, let alone for a variety of household types and locations. As a result, Generation has had to develop our own living wage benchmarks for the countries in which we operate, sending local colleagues to gather the prices of goods like food, housing, and utilities and then combine this data with publicly available sources. If a freely available and robust source for global living wages existed, it would be a game-changer for all organizations operating in our field to understand how their graduates fare in comparison.

THE PATH FORWARD

How might nonprofits and sectoral stakeholders gather more and better durability data? It starts with being willing to have the hard conversations and agreeing upon a universal standard that we collectively believe will better inform our programmatic decision-making. Because we insisted on data-gathering from the outset, Generation currently holds 40 million data points that track the learner life cycle, from application to five years post-graduation. But we can't do it alone. We would welcome a debate within our field about what combination of metrics that track employment status, job quality, wages, career growth, savings, living wage trajectory, and personal well-being would constitute the most important benchmarks for measuring the long-term impact of our collective efforts.

Governments and philanthropies can accelerate this journey by making durability a priority. Of course, grantees can and should have a wide array of delivery models and theories of change. But tracking impact against a universal durability data standard would not only be illuminating for funders, it would also generate insights for practitioners.

Success won't come quickly and will doubtless require numerous experiments in every sector to assess what's both doable and valuable. The charter school movement in the United States shows that a data trajectory is possible. While data-gathering initially focused on enrollment and performance (relative to public school district peers), it has, over time, expanded to high school graduation rates, college acceptance rates, and college graduation. Now some segments of the field are even pushing data-gathering to include income earned in the first job post-graduation.

Whatever sector we work in, we are all pushing for change that improves individual well-being and addresses massive inequities—and we want that change to stick. We want durability. And the only sure way to know whether our program impact matches our aspirations is to roll up our sleeves and, with humility and patience, commit to following the durability path wherever it may lead.

Mona Mourshed is the founding CEO of Generation: You Employed, a global employment nonprofit network that trains and places adult learners into careers that are otherwise inaccessible.

TIME FOR A THREE-LEGGED MEASUREMENT STOOL

Going beyond traditional monitoring and evaluation to focus on feedback can lead to new innovations in the social sector.

BY FAY TWERSKY

People have framed the conversation about measurement in the social sector in terms of monitoring and evaluation for decades. They shorthand it as "M and E" and serve it up as a generic, two-dimensional description for measuring nonprofit performance. *Monitoring* is the routine data collection and analysis conducted by an organization about its own activities, while *evaluation* typically means the kind of data collection and analysis conducted by an independent third party.

In many respects, these two complementary parts of measurement have matured and strengthened over time. Aided by technology, monitoring has developed to collect information about who is being served and with what level of frequency and intensity, and even to track short-term outcomes—all of which can inform decision-making. And the evaluation field is now more nuanced, with new approaches to answer a wide array of questions about outcomes, impact, and the factors that enable or inhibit change.

But, for all their advancement, these two building blocks are insufficient. We need a third leg of the nonprofit measurement stool to achieve more balance: *feedback*. Distinctly focused on the customer or constituent experience, feedback involves systematically soliciting, listening to, and responding to the experiences of nonprofit (or government direct-service-provider) participants and customers about their perceptions of a service or product. By listening to customers' experiences, preferences, and ideas, we can gain unique insights that will help improve the quality and effectiveness of social programs.

INNOVATION, REVELATION, AND AMPLIFICATION

Certain organizations are already leading the way in using feedback. Many have embraced customer perspectives as a crucial component of their work to source innovation, to surface hidden problems, or simply to amplify marginalized voices in our typical systems of service delivery. Examples of these three advantages underscore how adopting feedback into the measurement process can benefit both the programs and their respective clients.

Sourcing innovation | Some organizations implement all three legs of the stool, such as Nurse-Family Partnership (NFP), a model evidence-based program that recently used feedback to question its assumptions about what its clients actually wanted. NFP began in 1977 as a research project in Elmira, New York, whose studies determined that when a nurse regularly visits with a first-time mother for two years, providing a range of support and information, the arrangement produces many benefits, such as better birth and early child outcomes, and improved parenting. The program has been externally evaluated for 40 years and expanded into 42 states and 6 tribal communities. But in 2015, then-new CEO Roxane White and new Chief Communications and Marketing Officer Benilda "Benny" Samuels determined

that even evidence-based programs needed periodic innovations to reach new mothers and retain participation. So NFP decided to participate in Listen for Good, a systematic feedback tool to ask mothers about their experiences with the program—from whether they would recommend it to other new mothers to what they saw as its strengths and improvable areas.

Some staff were skeptical that these women would want to participate. They worried that the mothers wouldn't want to use their data plan minutes to respond to the Listen for Good survey via text message. But when staff sent the feedback survey to 10,000 recipients, they received almost 1,000 responses in 20 minutes. The first thing they noticed was just how much the mothers appreciated the invitation; they saw it as a sign of respect. While they provided positive feedback about the program overall, they also had innovative ideas for improvement, such as connecting the participating mothers with one another, not just with the NFP staff; creating an app for NFP's print materials; and, counter to the staff's expectation, asking to be able to communicate with the nurses both via text and in person.

These recommendations led to innovations now being tried at NFP, including creating a new feedback team that is not only engaging mothers but also inviting feedback from staff, volunteers, and partner organizations. NFP, the gold-standard, evidence-based program, has integrated feedback as a third leg of its measurement stool to unlock new insights and drive continuous improvement.

Surfacing hidden problems | The Second Harvest Food Bank serves millions of people each year in Silicon Valley but never systematically solicited feedback from its customers until 2016. From its first feedback efforts in 10 locations, Second Harvest learned that customers from different cultural communities were having vastly different experiences with food and service at the food banks—white and Latino clients were markedly more satisfied than Asian clients. This insight led Second Harvest to experiment with more culturally sensitive approaches to its work, including new-volunteer recruitment and training, food choices consistent with traditional Asian diets, and even a new location for food pickup, to better serve the Asian community.

Second Harvest, like many nonprofits, will likely never invest significantly in an expensive third-party evaluation, but rigorous systematic feedback has bolstered its understanding of client experiences and preferences. If the organization can improve clients' experiences, it will be better positioned to accomplish its mission to reduce hunger in all local communities.

Giving voice to those who are least heard | Epiphany Community Health Outreach Services (ECHOS) is a nonprofit ministry of the Episcopal Diocese of Texas that provides health and social services to the growing population of immigrants and refugees in Houston, Texas. ECHOS provides a range of critical safety-net services, including English language classes and, more recently, Hurricane Harvey relief services. Through its first efforts at customer feedback, ECHOS learned that clients were waiting excessive amounts of time, which consequently made ECHOS' support difficult to access.

ECHOS' staff realized that this experience contradicted their intention to treat immigrants and refugees with respect. Consequently, ECHOS is changing its registration process so that clients no longer have to stand in line to be received. It is also instituting expanded hours and making workflow improvements to increase efficiency. In fairly short order, the organization has transformed how it manages the flow of people, in order to promote a more positive and respectful customer experience.

ITERATIONS ON FEEDBACK

When I began my career in applied research 30 years ago, I was taught that client-satisfaction surveys were useless. They were seen as "lite," in contrast with "hard," outcomes. Because of the power differential between nonprofits and their clients, evaluators assumed that satisfaction measures would always be positive and therefore not meaningful.

It's a new day now. Throughout the social sector is a growing recognition of the importance of being human centered—that is, of putting the people we seek to benefit at the center of problem solving. That human-centered design principle should also apply to nonprofit measurement.

Many funders are already interested in connecting more with the communities they aim to serve, as well as looking for new measurement tools. In fact, a recent study by the Center for Effective Philanthropy found that foundation CEOs believe that listening more to the people they hope to help is essential to their success. The Fund for Shared Insight, the philanthropic collaborative that has been the driving force behind Listen for Good and other feedback efforts, has grown rapidly in the past four years, from 6 participating funders to 78 co-funders and counting.

US-based funders, such as the Plough Foundation in Memphis, Tennessee, have found feedback to be a powerful tool for their grantees. "The organization had never asked [its] populations what they wanted," says Diane Rudner, the Plough Foundation board chair. She describes the learning from feedback, particularly from an organization serving people with developmental disabilities, as "so valuable … it's amazing!"

Internationally focused Omidyar Network partnered with the Acumen Fund on "lean data sprints," in which they gathered feedback from approximately 30,000 customers from 68 Omidyar investees across 18 countries. Like Listen for Good, the Acumen tool uses the Net Promoter system, which involves a calculation of customer experience scores. Based on this first effort to be customer-centric in its measurement, Omidyar and its investee organizations generated actionable insights about each relevant sector—such as independent media, education, and financial inclusion—and each organization's perceived strengths and weaknesses.

The Omidyar Network still relies on traditional impact evaluation when it can, and on elaborate dashboards to monitor the progress of the organizations and businesses it supports. But now that the company has added a third dimension, *customer feedback*, to its measurement stool, it's positioned for customer experience to drive improvement.

Let me be clear: I am not arguing against monitoring or evaluation. They are both important tools. Evaluation helps us to gain a deep understanding of what works and why, and monitoring helps us track our progress and provides useful signs for course correcting. But not every organization can invest equally in each leg. The advantage of feedback, when properly integrated, is that it is both information-rich and affordable. The insights, ideas, and preferences of our ultimate beneficiaries can unlock new possibilities for operational improvements, programmatic innovation, and more respectful engagement.

As a mentor of mine advised, "Let not the abuse of a thing be an argument against its proper use." It's time to stop denigrating satisfaction surveys and unleash the power of feedback in new ways. Let's strengthen our measurement tools to be reliable, comparative, and simple to use, with both quantitative inputs and qualitative comments. And let's start listening to gain insight, to improve, and to innovate.

Fay Twersky is president and director of The Arthur M. Blank Family Foundation.

The Institute of Philanthropy

was established in September 2023 through a strategic seed grant of HK$5 billion ($640 million) from The Hong Kong Jockey Club and its Charities Trust. Established as an independent "think-fund-do" tank for China and Asia, the IoP is dedicated to promoting philanthropic thought leadership and enhancing sector capabilities at local, regional, and global levels in collaboration with fellow funders. It seeks to provide an Asia-based platform bringing global stakeholders together to promote the betterment of societies everywhere.

公益慈善研究院
Institute of Philanthropy