# Priorities for a New Decade: Making (More) Social Programs Work (Better)

Nadya K. Shmavonian 09 June 2011

Submitted by Nadya K. Shmavonian on behalf of Public Private Ventures.

## The Challenge

Thirty years ago, the social sector was rich with innovative models and services but starved for hard evidence that any of them would actually work. In those days, when Public/Private Ventures (P/PV) was just getting started, there were virtually no organizations equipped to find and evaluate promising programs, nor was there any consensus about the best way to measure program effectiveness.

Much has changed in three decades. Today a whole industry stands ready and willing to evaluate any new, old, promising or faltering program. Public and private funders increasingly ask for evidence of effectiveness from the programs they support. Internet-based tools designed to enhance program effectiveness abound. One evaluation method in particular—the Randomized Controlled Trial (RCT)—is now generally regarded as the most

scientifically rigorous and accepted way to evaluate social programs, and its use is spreading.[1] The wealth of new evaluators and research methods has produced volumes of information and analysis that were scarcely imaginable three decades ago.

Yet even amid this flood of data, the promise of using information gleaned from evaluations to improve programs remains elusive. Despite increased pressure to report on outcomes, most nonprofit organizations do not have the resources to collect and analyze data in ways that could help them boost their performance. Evaluation is often something that is done to programs by funders who hire external evaluators. Many evaluations fail to yield information that is of immediate, practical value to programs, including information about how the program could be spread to new settings. Funders themselves (both public and private) are not consistently asking for the right kind of evidence, at the right time, or for the right programs.

Nonprofits that have tried to build capacity for collecting evidence and using it to inform progress have found it difficult to raise money for these purposes. The emphasis that charity watchdogs place on administrative spending has made it even harder for organizations to invest in collecting, analyzing and using data.

Where does this leave us? As matters now stand, there is considerable interest in sophisticated methods of judging

program effectiveness. While this work is important, too little attention is being paid to actually making programs more effective—that is, improving the on-the-ground practices and implementation of social programs. And there is a very real risk that the current evidence-based trend will quash organizations whose work has not yet been or cannot be conveniently evaluated.

Too many innovations happening at the program level still go unnoticed by researchers, funders and policymakers. Too many evaluations examine programs that are poorly implemented, too young or inadequately funded—which produces inevitably disappointing results, even though the models behind the programs might work if given more time and better execution. Too few programs are based on clear, evidence-based theories about how they will accomplish their goals. And too few nonprofits have the capacity or receive the technical assistance needed to better use their data to strengthen their programs. In the rare instances when programs are proven to work, too little is known about how to successfully scale up those programs without diluting their impact.

Yet effective social programs are needed now more than ever. The outlook for people living in our country's poorest neighborhoods is bleak. Unemployment is at the highest rate ever in postwar America, a problem that is likely to persist for at least a decade, and unemployment for 16–24 year-olds is particularly acute. Unemployment and incarceration rates among African American and Latino

youth are significantly worse than for youth overall. These problems are compounded by an ever-widening income gap and shrinking job opportunities for low-skilled adults.

Confronting problems on this scale calls not just for tested models and proven services. It calls for organizations that can reliably deliver those models and services to the huge population that needs them. The evaluation field has produced a substantial toolkit for gathering information and assessing programs. But we are now faced with the far grittier issue of how to actually improve program effectiveness—and do so at a scale that stands any chance of ameliorating the grave social problems that continue to plague our country.

1. ^ A study is “randomized” when a population of eligible participants is assigned at random either to take part in an experimental program or to belong to a “control” group that does not receive the program’s services. The evaluation consists of comparing the experiences of the two groups to determine whether participation in the program led to materially different outcomes. The extent of the difference between the two groups is then deemed to be the “impact” of the evaluated program.

## A New Agenda: P/PV Case Study

# A New Agenda

The good news is that it can be done. We can learn from the past 30 years of evaluation experience. Much as P/PV and others have invested in and developed credible technologies and capacity for evaluation, we must build on this work so that on-the-ground program quality and effectiveness make equivalent progress over the next decade.

If there has ever been a time to advance this new agenda, it is now. The drive to identify evidence-based programs has intensified as resources have become more scarce. Private philanthropy continues to be constrained by the sluggish economy and its lingering effect on foundation spending. Gaping budget deficits at all levels of government have increased pressure to fund only programs with evidence of effectiveness. The sense of urgency to make resource allocation decisions wisely could not be greater—and for those of us in the business of evaluating social programs, it borders on moral imperative.

Yet if the past is any predictor, simply imposing more impact evaluations, data collection systems or requests for evidence is unlikely to produce significantly more programs that we are sure work. Effective programs rarely arise fully formed. They are nurtured and grown into being effective. Change must happen through the leadership of skilled practitioners who make sound use of evaluative information to test and improve programs and share their experience—which in turn can inform and shape relevant

policies.

Evaluations should be designed in ways that invite practitioners to make use of the results and to adopt solid practices based on evidence. We need to recognize the role of motivated program leaders at the center of evaluation efforts, to ensure that these efforts advance program theory and practice, rather than merely fulfilling a funder request. Research tells us that organizations and their leaders need to own and trust information in order to use it. Partnership on the ground between practitioners and evaluators, along with the long-term support of committed public and private funders, is indispensable if the goal is to deliver evaluations that actually improve program quality and effectiveness.
Reaching that goal will require at least two major steps:

1. The field needs clearer guidelines on how evaluation can meet the particular needs and contexts of different kinds of programs.

   Too many programs lack clarity about the theories that underlie their work. Too many have transformative goals and inspirational leaders but a limited capacity to realize those goals and inspirations (or especially to take their programs to a larger scale). At various stages of the cycle of innovation, programs require distinct approaches to program assessment. Practitioners, funders and policymakers need guidance and discipline in

applying their evaluation resources to each kind of organization and stage of development—and need to manage their expectations according to what can realistically be learned at each stage. We are faced with several important challenges that could benefit from greater clarity, standards and guidance:

The first is to **promote a menu of credible evaluation alternatives that can be used when an RCT or other impact methodology is not suitable**. There are many programs that are simply not appropriate for random assignment. For example: those that are too small or too new, those that are struggling with implementation challenges, programs that don't turn any applicants away and thus can't create a control group, or programs that provide broadly enriching experiences for young people (visiting museums, playing sports) rather than attempting to make a distinct measurable impact with a precisely defined intervention. Learning about what constitutes program effectiveness in each case is not just a matter of gauging causality, as RCTs do, nor of collecting data on outcomes, which is a frequent technique for programs that can't afford (or aren't ready for) an RCT. Good implementation research that gathers information about the how, who and what of a program's day-to-day execution is also vitally important, particularly for establishing the replicable mechanics of good practice.

The second challenge in making a better fit between evaluation and distinct types of programs is **to improve the use of RCTs**. Here, the simplest statement of the problem is the maxim that if all you have is a hammer, everything looks like a nail. While RCTs are a vital mechanism for assessing program impacts, they are a means to a certain kind of knowledge under certain circumstances. They are not an end in themselves. RCTs and other rigorous impact studies can easily be imposed inappropriately (and at significant cost), leading to unfair and unhelpful generalizations about program effectiveness.

A study published in the Archives of Pediatrics and Adolescent Medicine in 2010 illustrates this point. The study examined the statewide implementation of the Nurse-Family Partnership in Pennsylvania, which P/PV helped to manage. The program model had already been proven effective in rigorous evaluations elsewhere, yet an early study conducted during its first three years in Pennsylvania seemed disappointing at first. Focusing on just one aspect of the program, the evaluation had found only a “muted” effect on teen pregnancy. A few years later, the same researchers returned and found a significant impact. What happened? The answer is that it took a little time and considerable effort to get the

implementation right. Once that happened, the program showed the same strong results it had in other places. “Successful implementation,” the researchers concluded, “likely aided program maturation … and reduced second pregnancies.[2]

In short, a useful impact evaluation is not just a matter of knowing how to design a good RCT; it depends just as much on knowing when the service is ready to be evaluated. RCTs and their counterparts can be more effectively applied when:

- They are focused selectively on programs for which they are best suited;
- They test impacts that are realistic, as indicated by a sound research-based logic model;
- They are accompanied by timely and robust implementation studies—research that shows what actually happens in the program, how it is operated and managed, and ultimately why it did or did not work; and
- Their findings are used to clarify theories of program effectiveness and to generate relevant and helpful lessons for practitioners and policymakers.

Furthermore, while RCTs have been used extensively to test overall program impacts, the method has not typically been employed to test the effectiveness of specific program practices—though it could be. Nonprofit organizations are eager for definitive

answers to questions about practice: Which ones really work, and which should be changed? Evaluators generally approach these questions in an exploratory way, but in some cases it is possible to answer them more rigorously—by experimentally manipulating different program components (for example, the length or intensity of a program, or the type of training and support provided to program staff). By combining this research with thoughtful cost/benefit analysis, we can determine if programs, funders and taxpayers are getting their money's worth for various practices.

A third way to improve the usefulness of evaluations for different kinds of programs is **to help more nonprofits use common systems of evaluative information at a reasonable cost**. The myriad data collection systems and reporting processes that have emerged in recent years have done little to help programs actually use data to improve their performance. Worse, they have created separate, fragmented systems of measurement that make it hard to compare one outcome with another, or to discern which ones actually constitute "good" performance. One alternative, well worth exploring, would be to develop common measures that can be used across similar fields of practice. Working with a broad selection of nonprofits and funders in a given field—and armed with past research—it should be

possible to reach agreement on some basic elements of program effectiveness and on standard ways for all organizations to measure those elements. This approach, known as “common measurement,” makes it possible to compare one organization’s progress with that of another, and to create basic benchmarks of quality.

In fact, P/PV’s six-year experience with The Benchmarking Project shows that even within a fragmented field such as workforce development, it is possible to develop approaches to measuring performance across a whole field of practice, with similar organizations measuring and comparing their outcomes.[3] The Benchmarking Project has enlisted more than 200 workforce development programs to share and compare their data and has created a forum where participants can exchange ideas and experiences and help shape the project’s future. The trove of information and the eager participation of so many frontline practitioners has in turn made it possible to formulate persuasive recommendations for funders and policymakers about how they can support program improvement across the workforce development field.

So far, common measurement has been tried only sporadically, usually by individual funders working only with their own grantees. While these efforts have

the potential to be extremely useful in building and improving whole fields of practice at a reasonable cost, they have not been sufficiently assessed to know which ones work best and why. As various approaches become better understood, public and private funders can help lead the charge to bring more comparability across the fields in which they work. One benefit could be to reduce the pressures on nonprofits to come up with more and more kinds of data to respond to the particular interests of multiple funders. It could instead encourage them to help create and adopt common benchmarks with which to interpret their own programs' performance and—importantly—to use those benchmarks to guide improvement.

Finally, another approach to making evaluations meet programs' real needs would be **to develop more rigorous standards and practices for scaling and replication**. When evaluations produce encouraging results, there is an understandable urge to push the promising program toward greater scale. But bigger isn't necessarily better for every model or organization, and there is considerable murkiness in the nonprofit sector about exactly what is meant by "scaling up."

Frequently, "scaling up" means replication. But previous efforts to replicate effective models have

been spotty at best. Consider, for example, the attempt in the 1990s to expand the Center for Employment Training, which had produced impressive results in its two original sites in San Jose, California. When the federal government attempted to take the program to 12 new sites nationwide, an RCT found that the replication sites significantly underperformed the original ones. One reason: Only 4 of the 12 new sites had managed to replicate the original model faithfully.[4]

The ideas of "scale" and "replication" can seem deceptively straightforward and mechanical. In reality, these seemingly simple terms tend to paper over a raft of challenges—organizational, technical, human, financial, and sometimes political—that can derail even the strongest programs if not carefully thought through and addressed. In the next decade, it is vital that we establish a better understanding of which methods of "scaling up" make sense in which contexts—and how to implement each method well.

Among other things, expansion or replication often requires that a carefully honed, well-tested model be adapted to the differing conditions of a new site. Yet there is little solid research to help organizations strike an appropriate balance between adhering to an original model and adapting to local circumstances. More information is needed about the level of

support required to help different kinds of organizations replicate a model successfully. And it is not clear what alternatives to "pure" replication—in which programs exactly copy a proven model's essential elements—are most likely to be effective. Can organizations modify existing programs—importing key practices or tools, based on others' research findings—and achieve similarly strong results? Practitioners need more evidence about the choices available to them and guidance to help them choose the best approaches to scale, given their particular challenges and opportunities.

2. Individual organizations and fields of practice should have the chance to demonstrate that they will use evaluative information for program improvement, if afforded the opportunity to do so.

   Too often, organizations are evaluated—a passive verb for an all-too-passive role—rather than joining as partners in the generation and use of evaluative information. Relying on research designed and conducted entirely by third parties, funders or evaluators may end up imposing program models with little attention to the strengths and weaknesses that implementing agencies bring to the table. The result is not only the loss of an important source of real-world experience in designing and interpreting research, but also a missed opportunity to instill a

sense of ownership among the people who will ultimately have to translate findings into action.

Service providers have made it abundantly clear that, given the chance, they will rise to the challenge of identifying, gathering and using evaluative information. Consider, for example, The Benchmarking Project described earlier. Its success lies in the fact that it was designed to use data that practitioners are already required to collect (often set by legislation or government regulations). Practitioners are able to submit data relatively easily and can identify how they are performing compared with other similar programs. As a result, they have joined the project enthusiastically, supplying and critiquing their own data and becoming active participants in a "learning community" (with online and in-person elements) that uses both the data collected and the experiences of participating organizations to discern effective program strategies.

Without a close working partnership between practitioners and evaluators, evaluations may reveal whether programs are meeting some a priori expectations, but not whether those expectations were the best ones, or how programs might accomplish even more, or how similar organizations might rise to a similar challenge. Yet those are the questions most funders and policymakers actually

want to answer. Doing so will require at least four essential steps:

First, **the experience of practitioners, researchers and funders needs to be a fundamental element in the design and testing of new program models**. Effective programs are not only grounded in theory and research; they tap practitioners' and funders' practical experience and sense of what works. To raise the odds of success, the exchange between practitioners, researchers and funders should involve frank ongoing discussions of what needs to be tested and why, and of the implementation challenges that must be overcome.

Second, **practitioners need to be involved in the design of evaluations and data collection systems**. These systems should be minimally burdensome for staff, building wherever possible on data they already collect. And the data generated should be of immediate practical value for program improvement, not just a means of determining success or failure after the fact. Systems that reflect the day-to-day realities of program management and implementation produce better data and are ultimately more useful to practitioners, funders and researchers.

Third, **practitioners, evaluators and funders need**

**to work together to develop new ways of assessing program models and organizational capacity**. When faced with disappointing results from a replication or evaluation, funders and evaluators alike tend to bemoan organizations' uneven capacity to implement new programs. Yet much remains to be learned about how program models and organizational capacities "meet." Evaluations should explicitly investigate the strategy being implemented, the organizational capacities to implement it and the interplay between the two. The field needs clarity about the basic organizational requisites for implementing or enlarging different kinds of programs. But developing new approaches to programmatic and organizational assessment will be valuable only if practitioners perceive the results as useful to them, as manageable in the course of their day-to-day operations, and as meaningful in the lives of the people they serve.

Finally, **funders should invest in translating evaluation findings into practical lessons for program leaders and practitioners**. In today's information-saturated world, program staff at all levels need to be able to sort through the constant barrage of data and understand quickly what research can tell them about effective program strategies. They need information that is fresh, easy to grasp, and reinforced by discussion with their

peers. It is not sufficient to write an evaluation report that is read once and filed away. The products of evaluation must be more creatively designed to meet practitioners' needs. Such products may include real-time "dashboard" data; workshops, training and peer-learning opportunities; and reports, guides and tools that are broadly disseminated to help practitioners improve their programs.

Overall, the emphasis for the next several years needs to be focused at least as much on helping practitioners understand, use and improve data within their programs as it has been, for the last three decades, on refining the techniques by which we evaluate social programs from the outside. The main purpose of those techniques, after all, has not been to enrich the field of data analysis. The value of 30 years of technical and analytical progress will be realized only when well tested programs—and the policies that govern and support them—actually improve lives and create opportunity, overcome disadvantage, and contribute to a society capable of responding effectively and accountably to its gravest social problems.

P/PV is a national nonprofit research organization that works to improve the lives of children, youth and families in high-poverty communities by making social programs more effective. For more information, please visit www.ppv.org.

**Ms. Shmavonian is President of Public/Private**

**Ventures. Prior to joining P/PV, she worked as an independent consultant, providing strategic direction and counsel to many private foundations and a broad array of local, national and international nonprofit organizations. She has extensive foundation management experience, most recently having served as vice president for strategy at the Rockefeller Foundation. Earlier in her career, she spent 12 years at The Pew Charitable Trusts, where she worked as executive vice president, following several years as a director of administration and a program officer in health and human services.**

1. ^ Rubin, David M., Amanda L. R. O'Reilly, Xianqun Luan, Dingwei Dai, A. Russell Localio, and Cindy W. Christian. 2010. "Variation in Pregnancy Outcomes Following Statewide Implementation of a Prenatal Home Visitation Program." Archives of Pediatrics & Adolescent Medicine.
2. ^ More information on The Benchmarking Project, including a 2010 report with interim recommendations for funders and policymakers, is at http://www.ppv.org/ppv/initiative.asp?section_id=26&initiative_id=36.
3. ^ Miller, Cynthia, Johannes M. Bos, Kristin E. Porter, Fannie M. Tseng, and Yasuyo Abe. The Challenge of Repeating Success in a Changing World: Final Report on the Center for Employment Training Replication Sites. New York, NY: MDRC, 2005.