

EXPANDING THE EVIDENCE
UNIVERSE:
Doing Better By Knowing More

Lisbeth B. Schorr

Frank Farrow

July 2011

For Discussion at the Harold Richman Public Policy Symposium

Foreword

This paper was developed for the **Harold Richman Public Policy Symposium**, the first in a series of forums honoring the memory of Harold Richman, who, with his friend and colleague Tom Joe, founded the Center for the Study of Social Policy in 1979 and chaired its board from 2000 until his death in 2009.

Harold made many contributions to the field of children's and family services. His distinguished career reflected his belief that actions to improve the lives of children and families require a rigorous understanding of research and a commitment to policy innovation. Harold began his life's work as a White House Fellow from 1965 to 1967. He was also the Hermon Dunlap Smith Professor at the University of Chicago and from 1969 to 1978, Dean of its School of Social Service Administration. Harold went on to found the Chapin Hall Center for Children at the University of Chicago in 1985.

This paper addresses issues and provides recommendations related to expanding the knowledge base necessary to improve outcomes for children, families and communities. We recognize that not everyone will agree with these ideas, and we put the paper forward in the spirit of furthering discussion, prompting additional debate, and contributing to the shared urgency for moving forward with innovative, purposeful and evidence-informed strategies that can improve life outcomes for children, strengthen families, and build healthy, safe and supportive communities.

We thank the many people whose ideas and experiences we draw upon in this paper. We particularly thank Anne Kubisch, Prudence Brown, Nan Stone and Tom Kelly for their astute review and critiques of drafts of the paper. We are indebted to Leila Fiester for her masterful editing and her unmatched ability to give clarity to the expression to ideas. We would also like to honor and thank the many people whose work is cited in the paper, and especially salute those leaders in government and philanthropy who are balancing the related needs for innovation and evidence as they do their work every day.

We hope that the symposium is a forum for sharing diverse ideas, reaching consensus and civilly debating differences, and renewing commitment to make progress in creating opportunities for all children, families and communities – all of which were markers of Harold Richman's work and life.

CONTENTS

FOREWORD	i
CONTENTS	ii
EXECUTIVE SUMMARY	iii
INTRODUCTION	1
THE TIME IS RIGHT TO EXPAND OUR DEFINITION OF CREDIBLE EVIDENCE.....	3
SECTION 1: COMBINE FINDINGS FROM RESEARCH, THEORY, PRACTICE, AND EVALUATION	6
SECTION 2: BECOME MORE STRATEGIC TO SUPPORT SUCCESSFUL IMPLEMENTATION AND SCALE-UP	13
Identifying Commonalities	14
Improving Replication and Scale-Up.....	15
Challenges to Replication and Scale-Up.....	18
SECTION 3: ADOPT A PRAGMATIC APPROACH TO ASSESSING COMPLEX INTERVENTIONS.....	23
Constraints on What is Considered Credible Evidence	23
Limitations of Current Evaluation Techniques.....	24
A Pragmatic Approach to Evaluating Complex Efforts	26
SECTION 4: CREATE AN EXPANDED LEARNING FRAMEWORK AND MANAGE TO RESULTS.....	34
SECTION 5: STRENGTHEN MEASUREMENT FOR ACCOUNTABILITY AND LEARNING	40
CONCLUSION.....	45
ENDNOTES	A-1

EXECUTIVE SUMMARY

Research and experience over the past two decades have provided more knowledge than ever before about what it takes to improve outcomes for disadvantaged children and families. But despite the nation’s expanded knowledge, we have not been successful in achieving significantly better outcomes at a magnitude that matches the need in critical areas such as healthy births, school readiness, school achievement, physical and mental health, and safe neighborhoods. Among the many reasons for this is one for which the levers for change are clearly visible: our failure to marshal the full extent of available knowledge, apply it to complex problems, and to generate new knowledge from the most ambitious efforts underway to address our toughest social problems.

In the hope of minimizing risks of squandering precious resources, leading public and philanthropic funders are constructing a framework for what is considered credible evidence that wisely raises the emphasis on using evidence to guide investments. However, we suggest that the boundaries which the prevailing framework draws around acceptable evidence too greatly limit the knowledge base available to policy makers, program designers, and evaluators.

Programs and practices that are proven through experimental methods are an important component of effective interventions, but to achieve significantly better outcomes on a larger scale, they are best seen as a take-off point rather than the final destination. Our commitment to ensuring that practices, policies, and strategies are “evidence-based” must be undiminished, but our definition of what counts as credible evidence must expand. Especially at this time of severe pressure to use scarce resources prudently, we must make use of all the knowledge we can muster—from multiple sources—to maximize the impacts of public and philanthropic investments.

Accordingly, we propose a five-part set of concrete actions that the philanthropic, public, non-profit, academic, business and entrepreneurial sectors can take to build a wider and deeper evidence base that we believe would contribute to substantially improved outcomes for disadvantaged children, families, and neighborhoods.

1: Combine findings from research, theory, practice, and evaluation to promote more informed decision-making

The idea that our knowledge about what works should come primarily from evaluations of a relatively small number of flagship programs does not take us far enough. These findings are the start of a knowledge base. The proven programs that exist today, even when scaled up, cannot achieve the magnitude of impact needed to change outcomes for the most disadvantaged children, families, and neighborhoods embedded within a host of broader federal, state and community activities.

To generate additional useful guidance on where governments, philanthropies, and local reformers should direct their efforts and resources in order to improve outcomes widely and substantially, we must use the combined findings from evaluation with findings from other research, theory, and practice.

2: Become more strategic to support successful implementation and scale-up

By looking at effective programs and strategies not just individually but also in clusters with similar goals, we can identify the common elements that contribute to success; these may turn out to be even more useful in helping communities know what to do as they adapt elements of proven programs to complex and evolving situations. Knowing the common factors of what works to achieve specified

results would enable providers, funders and community leaders to improve on past practices and to engage in continuous cycles of innovating, testing, retesting, and reassessing to make sure that implementation is optimal and that outcomes are achieved in the face of changing client characteristics, evolving economic and social environments, and new learning.

Syntheses of knowledge about *what has worked*, and *how*, will help to make interventions more effective and implementation stronger and will expand opportunities for successful scale-up.

The next generation of efforts to achieve transformative outcomes is likely to involve not only replication of individual model programs but also the more difficult task of building innovatively on effective current programs. This will involve strategically adding missing pieces and linking effective programs and programmatic strategies—to each other, to reformed systems, and to an “infrastructure for change” with the capacity to monitor, improve, and sustain them at high quality—all in the interests of addressing the needs of more people more effectively and improving results for whole populations.

The “wicked” problems that face us today tend to be caused by such complex forces that their course cannot be changed by isolated interventions. They require multiple stakeholders working together, over many years, with a shared commitment to common results, so that the resources and authority necessary to bring about the needed changes can be mobilized and successfully applied.

3: Adopt a pragmatic approach to assessing complex interventions

The solutions to today’s complex problems reach across multiple sectors and involve actions at many levels; they require a common agenda, a shared commitment to results by multiple stakeholders, regular measurement and feedback about progress, continuous communication and mutually reinforcing activities among all participants. These efforts can also contribute to the next generation of complex solutions if we mine them for all the learning and knowledge they can produce.

The randomized control trial is a powerful research design for some purposes. It can establish the efficacy of selected components of practice, as has been shown by its use in the medical field and its application to interventions that are conceptually neat, with a clear causal relationship to the outcome of interest. However, when causal connections are more diffuse, intertwined, and otherwise difficult to establish, we need not give up on assessing effectiveness. Rather, we must agree that the value of many kinds of interventions can be assessed, weighed, understood, and acted upon without having to be *proven* through experimental methods.

It is possible to put together rigorously developed data, even about complex interventions, that can lead to informed judgments about which interventions are most likely to be effective, which are probably less effective, and how to support the continuous improvement of the former.

We identify six elements of a pragmatic approach to evaluation of complex interventions: (1) beginning with a results framework; (2) using strong theory to connect activities to results; (3) being prepared to compare results, recognizing that it may not be possible to find a perfect comparison group that would *prove* causality; (4) using multiple evaluation methods to harvest the learning from complex interventions, matching these with the multiple purposes of evaluation, the nature of the intervention, and the stages of implementation; (5) using randomized experimental evaluation designs when that is feasible and when the focus is on determining with certainty whether or not a program or a component of a complex intervention, focused on a specific outcome, is achieving desired results; and (6) using non-experimental evaluation designs to assess and understand whether, how, and why the intervention achieved the desired results when causal connections between interventions and outcomes are more diffuse.

4: Create an expanded learning framework and manage to results

The idea that nothing is worth knowing unless you know it for certain has its place, but not when applied to complex social programs and policies.¹ We can learn so much, including about program effectiveness, without insisting on absolute proof.

Valuable “real-time” learning can be generated from complex interventions as part of the day-to-day management of the work by developing a results framework, tracking progress toward those results, and using the data to continuously shape, drive, and improve efforts. When the process of managing to results, and the learning that accompanies it, is adopted by many partners and adhered to across multiple service systems, it becomes a method of achieving truly ambitious improvements in child and family well-being.

Shifting to a more results-oriented, comprehensive, and integrative accountability system will require investment in: (1) ongoing community capacity to gather, analyze, and process data needed for decisions across systems and sectors; (2) people with the skills and expertise to staff processes by which multiple partners review data and experience, learn from it, and chart their future course; (3) the capacity of neighborhood residents to be influential leaders in this process; and (4) links to citywide, regional, state and federal decision-makers who control many of the resources needed to achieve results.

5: Strengthen measurement for accountability and learning

The paucity of good measurement tools is a formidable barrier to maintaining accountability, managing by results, continuously improving quality, and assessing impact in complex initiatives. If high-quality, widely accepted, readily understood, user-friendly and reliable measures and indicators are to be available where they are most needed, philanthropy (initially) and the public sector (eventually) must become more intentional about investment in developing appropriate data sources, indicators, and measures by:

- Developing appropriate measures for smaller geographic units, because few indicators are uniformly available to cities, neighborhoods, and other small areas where place-based reforms operate;
- Developing metrics to capture all critical areas of work, because what gets measured gets done, striving for as much clarity about what to measure as about how to measure;
- Creating appropriate interim measures and helping funders and political leaders understand why they should attend to incremental signs of progress that predict long-term results; and
- Helping all stakeholders to emphasize and work with shared results, contributions, accountability, and measurement frameworks.

¹This is the idea that the late MIT organizational theorist Don Schön described as “epistemological nihilism in public affairs,” the view that nothing can be known because the certainty we demand is unattainable.

If the pragmatic, inclusive approach to evidence advocated in this paper is to take root and flourish, we believe that public and private funders must take the following steps:

Step 1: Support knowledge collection, analyses, and evidence syntheses that yield a more complete body of evidence.

Step 2: Ensure that state- and community-level initiatives can generate rigorous new evidence

Step 3: Accelerate the development of the tools and capacities that will help local communities generate new knowledge at greater scale.

Step 4: Working with the evaluation community, continue to expand the menu of available evaluative techniques that can be matched to different types of interventions and different needs to know.

The debate about criteria for credible evidence is neither academic nor trivial. How we as a nation deal with issues of evidence will shape the nature of social innovation, programs, and policies—what is and what is not allowed, promoted, and incentivized—for years to come.

Too much potential for innovation, and for improved outcomes, will be lost if we continue to define credible evidence too narrowly. Effectively addressing poverty, inadequate education, joblessness, and years of disinvestment in low-income communities will require using all the evidence we now have *and* an aggressive, rigorous and inclusive approach to gathering new evidence about the nuanced and powerful strategies for change that are emerging.

We hope that this paper contributes new ideas to the discussion and moves us toward common ground on all the ways we can use today's ever-expanding knowledge, and continue to generate more.

INTRODUCTION

Thanks to the last two decades of research and experience, we now know more than ever before about what it takes to improve outcomes for disadvantaged children and families—whether one is talking about healthy births; school readiness; school

achievement; robust physical, social, and mental health; or neighborhood support and safety. Unfortunately, our expanded knowledge has not yet led to significantly better outcomes of a magnitude that matches the need. One reason is the daunting nature of the problems themselves, which include persistent poverty, lack of educational achievement for too many children, and the mismatch between workforce demands and the skills of low-income parents. Other reasons include severe budget constraints resulting from a weak economy, distrust of the role and effectiveness of government in tackling tough problems, and limited capacity to implement solutions.

This paper suggests another factor that contributes to the gap between knowledge and accomplishment—one for which the levers for change are clearly visible. Too often, when policy makers and philanthropists attack complex problems, those efforts do not draw upon the full extent of available knowledge. And they are only rarely able to generate new evidence and new knowledge commensurate with the innovations underway.

This shortfall has multiple causes. In the past, support for the evaluation of ambitious and complex initiatives (involving interacting programs, systems reform, and/or community change, for example) and for the infrastructure for learning has not been as robust as for the innovations themselves. The resultant learning has been sparser than it could have been. And formidable difficulties stand in the way of disseminating the knowledge that *is* available. Despite accelerated attention to best practices and efforts to share lessons learned, many attempts at innovation begin without being fully informed by a deep, practical, and useful knowledge base about what has gone before. Underlying these problems are fundamental factors that prevent us from using the full range of available knowledge and from adding to that knowledge base systematically:

Too often, when our public and philanthropic systems attack complex problems, those efforts do not use the full extent of available knowledge. And they only rarely generate new evidence and new knowledge commensurate with the innovations underway.

- **Our framework for what we consider credible evidence is constructed too narrowly.** Leading public and philanthropic funders are wisely emphasizing using evidence to guide investments. However, we suggest that the boundaries which the prevailing framework draws around acceptable evidence can limit the knowledge base available to policy makers, program designers and evaluators. We focus more on evidence that is generated from some sources than others and lack ways to assemble and use a complete knowledge base. For example, impressed by the gains that one type of experimental approach (randomly controlled trials) has produced in the physical, biological, and medical spheres, we've applied a similarly scientific approach to testing, designing, selecting, and learning from programs and policies aimed at curing social, educational and community ills. This has generated much useful knowledge—especially about programmatic interventions—but this source of knowledge should not be seen as the *only* reliable source of evidence, but one of many. We need to assemble knowledge about more complicated community conditions and capacities, new findings from research and practice, and new opportunities that are equally essential to achieving good outcomes.

- **We are not using the full range of methodologies needed to learn from innovative and complex approaches to solving long-standing problems.** Although strategies to tackle our toughest social problems are being implemented, we have not invented or installed the techniques that would allow us to reap commensurate learning from these investments. We need more responsive, sensitive, and cost-effective ways of learning in real time from current systems reform and community change initiatives. What’s needed goes beyond inventing new evaluation techniques (although those are necessary as well). Rather, it is a matter of equipping state and local innovators with the tools, measures, and accountability structures that can produce real-time learning. Doing this will have the double bottom line of (1) empowering the people who implement change to improve their strategies, based on what they are learning; and (2) ensuring that new learning becomes available more systematically to the field.
- **We have not solved the problem of using accumulated knowledge and experience to expand and scale up interventions, strategies, and initiatives that have improved outcomes for children, families, and/or communities.** This problem is closely related to the two above. Too often, attempts to move toward scale have focused primarily on replicating program models rather than also tackling the challenge of identifying the core principles and characteristics that make a strategy work, adopting those principles and characteristics with fidelity, and then adapting the successful intervention to fit the context of new circumstances. Gains have been made as people pay more attention to theories of replication and scale, but careful attention is still required if we are to grow well-known “branded programs” in a way that nests them within systems and community context so that significant improvements are achieved in outcomes for children and families at risk.

Put simply, to strengthen and accelerate our capacity to make major improvements on a large scale for the children and families most at risk of poor outcomes, we must take an inclusive approach to how we define, collect, assess, and use evidence. We should adapt our ways of knowing and learning to a social terrain in which organizational climate, neighborhood norms, context, infrastructure, and the “connective tissue” of community capacity may be as important as the more tangible and circumscribed aspects of social reform efforts that we have been more used to dealing with.

Our commitment to ensuring that practice, policies, and strategies that are publicly or philanthropically funded will be evidence-based or evidence-informed must not diminish. But our definition of what counts as credible evidence when judging what is worth funding or scaling up should be expanded to allow for continuing improvement and innovation. In addition, ways should be found to generate usable evidence, more systematically, from current initiatives in order to strengthen resource allocation, program design, and policy making—and ultimately to assure that more children, families, and communities achieve outcomes that are powerful, enduring, and cost-effective over the long term.

Our commitment to ensuring that practice, policies, and strategies that are publicly or philanthropically funded will be evidence-based or evidence-informed must not diminish. But our definition of what counts as credible evidence when judging what is worth funding or scaling up should be expanded to allow for continuing improvement and innovation.

THE TIME IS RIGHT TO EXPAND OUR DEFINITION OF CREDIBLE EVIDENCE

This is an opportune time to reassess what constitutes strong and credible evidence and to change how we obtain actionable information about complex social reforms.

An evidence base that would result in substantially improved outcomes for disadvantaged children, families, and neighborhoods would consist of:

- Findings from research and theory about what children need for optimal development
- Evidence from programs that achieve results and provide children and families with what they need
- Implementation factors and community capacities, connections, and infrastructure needed so communities can provide children and families with what they need
- Common factors of effective programs and strategies that achieve results
- Findings about the effects of complex interventions, based on multiple methods of evaluation as well as performance measurement using a results framework

Federally, the stakes have never been higher in terms of ensuring that new initiatives and innovations are effective and generate new learning to inform future reforms. The Obama Administration brought a fresh awareness of unsolved social problems and a renewed ambition to tackle them, emphasizing social entrepreneurship and innovation. The President consistently lauds innovation,^B and during his administration's first years federal agencies proposed new initiatives in rapid succession, triggering corresponding actions at the state and local levels. Given the rampant skepticism in this country about government programs, coupled with the dismal economic climate and political pressures to cut public spending, it seemed reasonable to minimize the risk that these public investments might fail to produce results. Thus in order to get some of these groundbreaking initiatives launched,^C rules and regulations were promulgated requiring that the bulk of funds go to interventions that had been previously proven through evaluations with experimental designs to produce strong or moderate evidence, even when that decision could be seen to be in tension with the determination to encourage innovative solutions to long-standing problems.

Those initial funding constraints, linked to a limited definition of what constitutes "evidence," helped launch the initiatives and communicated well the Administration's focus on evidence. Once established, however, we believe these initiatives and future ones could make the greatest strides toward solving urgent social problems if they take a more inclusive approach to what constitutes useful and credible evidence and if they install the measures, accountability approach, and learning infrastructure needed to generate new knowledge. Equally important is how this more inclusive approach could apply to large federal programs already on the books. It may well be that in the current climate of fiscal retrenchment the opportunities for innovation, improved performance, and greater cost-effectiveness are at least as

^B Several of the new initiatives (including two discussed in this paper, the Department of Education's Investing in Innovation Fund and the Social Innovation Fund) had 'innovation' in their titles. In his 2011 State of the Union message, the President identified innovation as the key to a thriving nation and called on Americans to "out-innovate ... the rest of the world." United States President Barak Obama. (January 25, 2011). State of the Union Address. <http://www.whitehouse.gov/thepressoffice>.

^C Some of the initiatives that seek to balance the need for innovation with a focus on proven successes are *Promise Neighborhoods* (Department of Education), *Choice Neighborhoods* (HUD), *Investing in Innovation Fund* (Department of Education), *Social Innovation Fund* (Corporation for National and Community Service, the White House), and Home Visitation Program of the Patient Protection and Affordable Care Act (Department of Health and Human Services).

large within existing funding streams^D as in the newer programs that may be most vulnerable to budget cutting.

Reassessments of how best to define, use, and generate evidence are equally urgent at the state and local levels, where governments are more financially pressed than ever before. Fierce budget battles create a context in which innovative efforts must justify themselves by providing evidence of what they accomplish. State and local leaders thus also face choices about what standards of evidence they will use to make decisions and how they will track progress and use learning. Their choices will help determine whether the mushrooming efforts to reform education, transform disinvested communities, and otherwise improve outcomes for children and families will produce unparalleled new knowledge or will be limited by using more narrow approaches to evidence that, by themselves, are not aligned enough with today's complex problems and tomorrow's most promising solutions. One way to address this dilemma is to emphasize that public resources should be directed to programs and solutions that are informed by the *best possible* evidence, including when appropriate (as opposed to when available) strong and moderate evidence (as currently defined by the federal government).

The philanthropic sector faces similar choices. The concept of “evidence-based” has taken hold within many foundations, and this emphasis—on sound reasons to justify investments—is for the good. But will philanthropic leaders approach their renewed commitment to evidence with a broad enough scope that allows the sector to fulfill its role as the source of truly path-breaking innovation? Will a more circumscribed definition of evidence mean that foundations' social interventions will be less bold than they've been in the past? Or will philanthropy recognize multiple kinds of evidence to make the case for new funding and invest in the tools and technology needed to generate the diverse knowledge essential for tackling tough social problems?

We believe that many of the tensions between taking risks with innovation and minimizing risks by relying on what is already proven might be resolved in the future with a more inclusive approach to what counts as evidence. “Evidence-based” does not have to mean experimental-based. When we draw on evidence from many kinds of research—not just program evaluations—and from theory and practice, even innovations can be evidence-based.

A commitment to assembling better and richer information about complex interventions means neither a retreat to fads, hunches, anecdotes, or good intentions nor a reluctance to identify and end support for the efforts that are ineffective. It does mean that the rapidly moving changes in the economic and fiscal landscape, on the frontiers of brain research and human development, and in what we are learning about the needs of the families, neighborhoods, and institutions that have been left behind require us to approach the challenge in bold new ways. As Peter Drucker has pointed out, “The greatest danger in times of turbulence is not the turbulence; it is to act with yesterday's logic.”¹

When we draw on evidence from many kinds of research and from theory and practice, even innovations can be evidence-based.

^D Such as the Elementary and Secondary Education Act (ESEA) and the Workforce Investment Act when these are re-authorized, the Child Care and Development Block Grant (CCDBG), and the Fostering Connections to Success and Increasing Adoptions Act.

We propose a five-part set of concrete actions that will move us away from yesterday's logic and toward a richer, more useful way to make decisions for allocating resources, formulating policies, and designing and assessing interventions:

- 1: Combine findings from research, theory, practice, and evaluation** so that decision makers can make more-informed decisions about allocating resources and selecting and designing programs, policies, and strategies;
- 2: Become more strategic to support successful implementation and scale-up**
- 3: Adopt a pragmatic approach to assessing complex interventions**, using multiple methods to generate richer evidence;
- 4: Create an expanded learning framework and manage to results**, and
- 5: Strengthen measurement for accountability and learning.**

Each remaining section of this paper explores one of those essential parts in depth.

This paper is a practical call for expanding our “theory of evidence” to create and use a richer knowledge base. We propose pragmatic actions that the philanthropic, public, non-profit, academic, business, and entrepreneurial sectors can take to build a wider, deeper pool of evidence and knowledge that can strengthen current efforts to improve outcomes for disadvantaged children, families, and neighborhoods. We do this out of determination not to settle for only modest gains from public and philanthropic efforts to improve results. We believe that taking a rigorous but more inclusive approach to evidence will allow us all to make more significant and cost-effective improvements, on a large scale, for the children most at risk of poor outcomes.

SECTION 1: COMBINE FINDINGS FROM RESEARCH, THEORY, PRACTICE, AND EVALUATION

Program evaluations are a good starting point in the quest for evidence about what works. They can suggest effective practices and approaches. They can provide raw material to explore to find common patterns across disciplines and domains. They provide clues to needed policy and systems change and to programs that can be replicated, adapted, or built upon. However, program evaluations cannot provide all the knowledge we need, because the proven programs that exist today—even when scaled up—are not powerful enough by themselves to change outcomes for the most disadvantaged children, families, and neighborhoods. When we limit ourselves to only the body of evidence that comes out of program evaluations, our knowledge is too circumscribed, too sparse, and too focused on discrete, isolated units of intervention.

The idea that our knowledge about what works should come exclusively from program evaluations is currently dominant, but is a thin reed to lean on for guidance about how to improve outcomes, especially for the people most at risk of long-term damage.

Lists of proven programs are now widely available. While user-friendly and helpful to answer important questions that administrators and funders face (e.g., “Is there evaluative data about specific programs I’m considering?”), their usefulness is constrained because: (1) they include only a scattering of programs; (2) they do not include programs that cannot be assessed with methods conventionally considered rigorous; (3) they rarely include interventions and strategies that go beyond the programmatic by changing systems, infrastructure, and norms; (4) they typically provide little or no information about factors that would allow people to implement a program with a community or population that differs from that of the original; and (5) they are silent about the many situations in which individual programs, no matter how good, have only limited effects—i.e., when it comes to populations experiencing clusters of risks to well-being.

In the early childhood field, for example, only two interventions, the Nurse Family Partnership and the Incredible Years, have been identified as “proven,” “model,” or “exemplary” by the majority of the 15 national registries of model programs surveyed by Child Trends, the highly respected national research organization.² Most programs that have been validated by experimental methods are circumscribed in one of two ways: the intervention is sufficiently narrow that it is possible to assess (for example, in the lists surveyed by Child Trends, one program “increased the likelihood of sunscreen being applied to [preschool] children” and one “increased the likelihood of usage of covers on electrical outlets”). More complex programs, on the other hand, tend to have been standardized in order to be assessed experimentally.

The Nurse Family Partnership, for example, is a remarkable case of a multi-part intervention that developed a model sufficiently standardized that it was possible to run randomized trials in three different sites. It is now operating effectively in 392 counties around the United States.³ The tradeoff this program had to make in order to maintain a single model that would look essentially the same from one site to another is that it would be self-contained. Were it to become part of a broader community-wide intervention, collaborating with community resources to work with mothers dealing with serious depression, domestic violence, or substance abuse, it would have to be adapted, because the connections with existing community resources would differ from one site to another. Similarly, if its target group were to include mothers who were high-risk even though they were not teenagers or first-time mothers, implementation would become more complex and would undoubtedly vary by community—which might

increase the impact on high-risk lives but also produce an intervention not “neat” enough to evaluate through randomized trials.

To generate useful guidance on where governments, philanthropies, and local reformers could direct their attention, efforts, and resources, we cannot stop with individual programs or settle for a one-dimensional hierarchy that ranks programs by how methodologically rigorous their evaluations were. The knowledge base that will lead to significantly better outcomes requires a more inclusive and systematic approach to accumulating, assessing, and arraying evidence and a broader definition of what constitutes credible evidence.

To take a very simple example, we know from research that children in foster care do better when they are in fewer out-of-home placements.⁴ We may not have the programmatic information about the nature of interventions that are routinely effective in bringing about fewer placements. But non-programmatic research can still guide action and resource allocation.

A more ambitious agenda depends on the work that cannot be done by individual programs. In order to tackle the toughest obstacles to improving outcomes, we need to be able to draw on the findings from evaluation combined with findings from other research, theory, practice, and, when it comes to the specifics of local implementation, local wisdom.

This broadened collection of evidence—a synthesis, if you will—would illuminate: what children and families need to reach specified results; practices, policies, programs, institutions, strategies, connections, and infrastructures that work or are promising to provide children and families with what they need; and essential components, adaptation options, and attributes that make these interventions work or make them promising.

Consider this possibility. What if, in addition to starting with a list of good programs, a community group that hoped to develop interventions to improve the health and well-being of children in its community could begin with the body of evidence assembled by a distinguished group of academics, such as the members of the National Forum on Early Childhood Policy and Programs and the National Scientific Council on the Developing Child? Operating under the auspices of the Harvard University Center on the Developing Child, these groups undertook an extensive review of “credible, peer-reviewed research” on “advances in neuroscience, molecular biology, genetics, and child development. They combined these findings with four decades of rigorous program evaluation data” and their own experience.^E

From this collection, the group produced a “Science-Based Framework for Early Childhood Policy” to establish the common ground on which all stakeholders could design effective policies and strategies.⁵ The Framework illuminates:

- The effects of early experiences on a child’s brain architecture, on the development of the body’s stress response systems, and on determining whether the foundation for future learning, behavior and health will be weak or strong.
- The significant reductions in chronic disease that could be achieved across the life course by decreasing the number and severity of early adverse experiences and by strengthening the protective relationships that help mitigate the harmful effects of toxic stress.

^E By integrating their own experience from observation and practice, these experts escaped the trap of what David Brooks calls an “amputated view of human nature,” relying only on “correlations that can be measured, appropriated and quantified.” Brooks, D. (March 7, 2011). “The New Humanism.” *The New York Times*.

- The public- and private-sector policies and programs that enhance the capacities of caregivers and communities in the multiple settings in which children live, learn, and play.
- The rapidly growing knowledge base that could strengthen investments in the early reduction of significant adversity.
- Policies that support the ability of parents, providers of early care and education, and other community members to interact positively with children in stable and stimulating environments in order to help create a sturdy foundation for later school achievement, economic productivity, and responsible citizenship.
- The strategies, now clearly identified, that are effective for children and families who are at special risk for poor outcomes.

These elements of what children, families, and communities need can inform efforts to identify the “pathways” or sequences of experiences,

interventions, and supports that are most likely to produce the better outcomes for children we seek. They can structure efforts to assess which relevant activities may already exist, which need to be improved, and which need to be supplemented. The success we want for children is much more likely to depend on a progression of effective supports for healthy development, academic achievement, and social integration than on any one intervention. It is the accumulation of these experiences that produce powerful and long-lasting results: young children entering school ready to learn and schools being ready to teach them, for example, or students reading proficiently by the end of third grade.

Building evidence for what it takes to accomplish good outcomes requires looking not only at the effectiveness of specific programs but mapping backward from the desired result to identify the interrelated and mutually reinforcing experiences, interventions, opportunities, and supports that collectively produce the result.

Building evidence for what it takes to accomplish these outcomes requires looking not only at the effectiveness of specific programs but mapping backward from the desired result to identify the interrelated and mutually reinforcing experiences, interventions, opportunities, and supports that collectively produce the result. From a research point of view, this is extraordinarily challenging. But from the perspective of the schools, health care providers, community organizations, parents, and many others trying to help entire populations of children, such guidance is essential.

Fig. 1 and Fig. 2 provide examples of ways that findings from evaluation might be combined with findings from research, theory, and practice to expand even further the reach and effectiveness of two federal initiatives, the Home Visitation Program of the Patient Protection and Affordable Care Act, and the Department of Education’s *Investing in Innovation (i3) Fund*. The section that follows looks more closely at one way to expand evidence of what works: identifying common factors of successful interventions that seek similar goals.

Fig. 1:
**COMBINING FINDINGS FROM EVALUATION, RESEARCH,
THEORY, AND PRACTICE
TO EXPAND THE REACH AND EFFECTIVENESS OF THE HOME VISITATION PROGRAM
OF THE PATIENT PROTECTION AND AFFORDABLE CARE ACT**

The Home Visitation Program, part of the 2010 federal health reform legislation, is intended to improve: maternal, prenatal, and infant health; child health and development; parenting related to child development outcomes; school readiness; and families' socioeconomic status. It also aims to reduce child abuse, neglect, and injuries.⁶

Discussions among the U.S. Department of Health and Human Services, Office of Management and Budget, and Congress originally focused on whether federal funds under this initiative should go only to replication of the Nurse Family Partnership (NFP), widely known as the most rigorously documented home visiting program model. NFP was considered a safe bet for scale-up because of its steady growth over 30 years and its positive results, measured through three randomized, controlled trials.⁷

Child development experts, however, expressed concerns that building a national initiative solely on the basis of a single model's target population and evidence from randomized trials provides little guidance on how to replicate the model at sufficient scale and scope to serve the national interest. They pointed out that NFP's unique focus on first-time parents would have left out 62 percent of newborns, and its exclusion of mothers who receive no prenatal care would have left out many of the highest-risk babies, including infants in the foster care system who are eight times more likely than other infants to have mothers who received no prenatal care.⁸

After extensive deliberation, the final federal guidance issued in February 2011 allows states to implement several home visitation approaches in addition to NFP, and to build on and adapt them. States also will be able to support public and nonprofit community-based organizations that are prepared to innovate by, as Harvard's Jack Shonkoff puts it, making proven programs their point of departure rather than their final destination.⁹

The use of knowledge bases that look more fully at what is known and understood about healthy child development and what strategies (not just programs) can reduce risks and build protective factors would further enrich implementation. Such knowledge could help program implementers and policymakers in the following ways:

Make a stronger cost-effectiveness argument for early intervention. Because home visiting programs typically start in pregnancy and continue for the first two years of an infant's life, states could apply the Harvard Center's Science-Based Framework's findings on the long-term effects of early intervention on a child's brain architecture, on the development of the body's stress-response systems, and on the foundations for future learning, behavior, and health—including reductions in chronic disease over the lifespan.

Claim long-term benefits for near-term achievements. Since the Science-Based Framework has established that the harmful effects of toxic stress (including impaired learning, maladaptive behavior, illness, disability, and a shortened life span) are averted when the number and severity of early adverse experiences decreases, and when protective relationships are strengthened, home visiting programs can focus on achieving these goals. As they do so, they can claim the long-term benefits without having to wait for the results to appear in the children now being served.

Target additional populations and incorporate better strategies for reaching them. Using the knowledge from the Science-Based Framework, states would be positioned not only to implement proven models of home visiting, but to adapt proven programs to include families struggling with substance

abuse, domestic violence, or serious mental health problems, who are currently excluded from many programs proven effective by experimental methods.^F Implementation could also be strengthened by adding strategies that the Science-Based Framework, drawing on a wider range of evidence than home visiting programs alone, has found effective for children and families who are at special risk for poor outcomes—such as treatment for depression for the mothers of young children experiencing abuse or neglect, or work-based income supplements for parents of children living in poverty. (Once they know what *can* be done, states would of course have to provide the supports, training, monitoring, and quality checks needed to ensure that these programs achieve the results they aim for.)

Incorporate additional functions and populations. Knowing from the Science-Based Framework that it is important to enhance the capacities of all caregivers, states could adapt proven programs by extending home visiting services to family, friend, and neighborhood caregivers (who care for 41 percent of low-income children under age 5 with employed mothers),¹⁰ as well as new mothers.

Moderate the effects of demographic risk factors. States could draw on evidence assembled by Child Trends and the Council of Chief State School Officers that children in families with low income, racial or ethnic minority status, a non-English home language, and low maternal education are significantly behind their peers' development at 9 months and even more so at 24 months. These disparities exist across cognitive, social, behavioral, and health outcomes. Since about half of these children are in some form of nonparental care, with most in home-based settings, the opportunities to intervene are considerable—especially since we know that high-quality nonparental care (both home- and center-based) has potential to moderate these effects at 24 months.¹¹

Influence policies and systems. Moving beyond programmatic interventions, the Science-Based Framework highlights the importance of policies that allow parents and providers of early care and education to interact positively with children in stable and stimulating environments. States could encourage and assist local community providers to (a) supplement home visitors by enlisting special expertise in mental health (especially maternal depression), family violence, and substance abuse, in the form of consultants or occasional partners on home visits; and (b) integrate home visiting into a broader place-based approach and a population focus, as partners in a larger universal system of support.¹²

Chapin Hall's Deborah Daro and Kenneth Dodge have made the case that improving the efficacy of home-visiting initiatives requires more than knowledge from evidence-based models.¹³ A knowledge base that combines findings from evaluation with a full spectrum of findings from other research, theory, and practice across systems and disciplines can help communities and states accomplish what no individual program can and can enhance states' ability to achieve the specified purposes of the home visiting law.

^F This is not a coincidence. The early home visiting programs set up by David Olds had to streamline their intervention and limit their target population in order to demonstrate their effectiveness in experimental studies by maintaining consistency over time and over a variety of service environments. Experimental evaluations must carefully limit the number of variables they include, sometimes circumscribing their study populations.

Fig. 2:
**COMBINING FINDINGS FROM EVALUATION, RESEARCH,
THEORY, AND PRACTICE: POTENTIAL IMPLICATIONS
FOR INITIATIVES LIKE THOSE FUNDED BY
THE *INVESTING IN INNOVATION FUND* (i3)**

The specified purposes of the Department of Education’s *Investing in Innovation* (i3) initiative are to accelerate the creation of an education sector that supports the rapid development and adoption of effective solutions and to expand the implementation of, and investment in, innovative and evidence-based practices, programs, and strategies that significantly:

- Improve K-12 achievement and close achievement gaps;
- Decrease dropout rates;
- Increase high school graduation rates; and
- Improve teacher and school leader effectiveness.¹⁴

i3, established in 2010, is of considerable current interest because it is the largest source of federal grant funding that defines the nature of the evidence required to make an applicant eligible for these funds. In authorizing grants to three types of initiatives, classified by the level of evidence of effectiveness they offer,^G it seeks to reconcile the tensions between striving for innovation (as the title of the fund suggests) and relying on traditional evidence of effectiveness.

The Department has struck a balance between the two. The majority of its first grants (30 out of 49 winning proposals) were awarded in response to proposals in the “Development” category (in which thresholds of evidence are more inclusive), while the bulk of funding has gone to applicants offering evidence from “well-designed and well-implemented experimental and quasi-experimental studies.”

The Department of Education has stated that it is planning to apply some of the principles it used in structuring i3 to other programs.^{15,H} Our conviction that an expanded view of evidence would strengthen efforts to improve outcomes leads us to ask whether the continuing emphasis on the rigor of the evaluation research, either already obtained or to be obtained in the future, is the best way to accomplish the purposes of such legislation. Too much emphasis on the use of this standard could minimize federal support for and discourage the development of more complex and less-standardized interventions—ones whose very complexity prevents them from obtaining evidence from more traditional evaluation methodologies. There are examples of federal education policy that are research-informed that lack the kind of strong or moderate evidence called for in i3. An example is federal policy around value-added teacher evaluations. There is solid research showing that high quality teachers make a significant

^G The largest grant category (Scale-up Grants) is focused on programs and practices with *strong evidence* from “well-designed and well-implemented (as defined in this notice) experimental and quasi-experimental studies.” A second, less well-funded category (Validation Grants) supports “programs that have good evidence of their impact from experimental and quasi-experimental studies,” or from correlational research with strong statistical controls. The category eligible for the smallest grants (Development Grants) goes to “support new and high-potential practices whose impact should be studied further.” U.S. Department of Education. (October 2009). *Investing in Innovation Fund*, <http://www2.ed.gov/programs/innovation/factsheet.html>.

^H Similar restrictions on evidence could also become part of the forthcoming extension of the Elementary and Secondary Education Act (ESEA). *Education Week* reported in August 2010 that the Senate Appropriations Committee called on the Department of Education to “encourage and support” states and districts to use their school reform funds only for interventions that would meet the evidence required for the two most stringent evidence standards in the federal i3, research grants—the “validation” and “scale-up” categories. Sparks, S.D. (August 26, 2010). “Senate Report Hints at a Definition for What Works.” *Education Week*.

difference in student achievement. However, there is not “strong or moderate” evidence that using value-added teacher assessment for accountability leads to high test scores. Nevertheless, this research-informed practice is supported by federal policy.

The priority that i3 assigns to the strictest research evidence also raises the question of whether the rigor of the documentation weighs more heavily than considerations about the impact of the intervention. Too great an emphasis on narrowly defined forms of evidence and on past programmatic success, without enough room for the development of more effective and innovative responses to unsolved problems, carries its own risks. We agree that the emphasis on strong and moderate levels of evidence, as defined by the Department, helped launch the initiative effectively. We suggest that the Department consider a broader definition of evidence in awarding a greater portion of its funds in the future in order to meet the initiative’s objectives of improving education outcomes. This would allow the Department to:

Dedicate a larger share of funds to multi-faceted interventions and strategies that cannot be readily evaluated by experimental methods because they cannot be standardized with all variables held constant, but that could have tremendous potential for success. With broader definitions of allowable evidence, the Department could assess interventions against *both* their evidence base *and* their likelihood for effectiveness, opening the door to a more complete range of innovations

Support the expansion and scale-up not only of proven practices, strategies, and programs but also those that are promising but require modifications and additions to become fully effective.

Take full advantage of the unprecedented public and political attention focused on education reform to push forward some truly transformational change at scale. The Department could advance and support ideas and strategies that may not yet have accumulated rigorous evaluative research but for which there is evidence of success in local practice from individual schools, clusters of schools, or in the experience of state systems or local school districts.

SECTION 2: BECOME MORE STRATEGIC TO SUPPORT SUCCESSFUL IMPLEMENTATION AND SCALE-UP

One way we could expand the evidence base about “what works” is by identifying the commonalities among successful interventions aimed at similar goals to shed light on how and why they achieve their impact. Once we possess and synthesize that information, decision makers can make better choices among alternative programs and strategies.¹⁶

Communities can then improve on past practices and engage in continuous cycles of innovating, testing, retesting, and reassessing achievements, even in the face of changing client characteristics, economic and social environments, and other variables.¹⁷ We also will be positioned to replicate good practices and programs even more strategically, making possible genuine scale-up—of strategies that work, rather than relying only on copying programs that may or may not work when brought to bear in different environments and with different populations.

This section explores the rationale for identifying commonalities and how research and experience, coupled with an expanded definition of evidence, can improve efforts to replicate and scale up what works.

As a nation, we are entitled to celebrate considerable accomplishments as we survey the landscape of skilled professionals, dedicated public servants, and committed volunteers who are increasing opportunities for those who have too few and transforming neighborhoods into welcoming places to live. Simultaneously, however, we must recognize that while the most effective interventions for children and youth living in poverty do produce positive outcomes, the magnitude and reach of their impacts are modest, and persistent disparities remain.

The Nurse Family Partnership (NFP) home visiting program is an example of one intervention that has been replicated and helped produce good results at a significant scale. However, we believe what has been learned through replicating NFP could have even greater impact if communities and program leaders were encouraged and had the tools to act on an additional and more inclusive body of credible evidence than what NFP produced through their program evaluation. NFP fields nurses to make home visits to low-income teenagers pregnant with their first child, until the baby is two. NFP has tested its model in three randomized trials (RCTs) with different demographics, beginning 34 years ago in Elmira, New York.

In all three RCTs, nurse-visited women had longer intervals between the births of their first and second children. Two of the three RCTs documented reductions in child abuse and neglect, reductions in health-care encounters for injuries, and improvements in language development and academic test scores among the subset of children born to “low-resource” mothers. Not found in any of the sites were several other positive changes that were sought: fewer behavior problems or fewer foster care placements among the children, or reduced substance abuse or psychological distress or incarceration among the mothers.

Nevertheless, the achievements were considerable, and because they were identified through randomized trials, NFP is widely considered the most rigorously proven early childhood intervention.¹ But to reap the benefits of being a “proven” program, this model, designed three decades ago, might be considered frozen in time since it has not been adapted to take account of the explosion of knowledge in the last two decades.

¹ NFP is the only early childhood program that has qualified for a “Top Tier” rating by the Coalition for Evidence-Based Policy.

If a proven program like NFP were also viewed as a take-off point for carefully considered adaptation, communities could build on its success by adding some of the following:

- Change intake requirements that exclude mothers who receive no prenatal care, since that exclusion leaves out many of the highest-risk babies, including infants in the foster care system who are eight times more likely than other infants to be born to mothers who received no prenatal care.
- Add partners that have worked successfully with new parents who are depressed or abusing drugs and alcohol, since we now know that toxic stress in early childhood leads to lifelong problems in learning, behavior, and physical and mental health, and that the two most common precipitants of toxic stress are parental substance abuse and postpartum depression.
- Add capacity to deal with crises arising from domestic violence or homelessness.
- Extend home visiting services to family, friend, and neighbor caregivers, who care for 41 percent of low-income children under age 5 with employed mothers.

However, communities that have enhanced their home visiting programs in these ways worry that they will be at a disadvantage in competing for funding, because their programs would no longer be viewed as proven.

The problem is that simply replicating intact program models is not a sufficient strategy for reaching enough vulnerable children and families with stronger interventions because the proven programs are often not validated for the wide range of children and family situations that exist. The danger is that by limiting implementation to only individual model programs that seem worth replicating in their proven form, we miss opportunities to expand, improve, and build on effective strategies to achieve greater impact.

Simply replicating “intact” program models is not a sufficient strategy for achieving greater impact. In fact, by getting stuck on identifying the individual model programs that seem worth replicating, we are failing to expand, improve and build on effective strategies.

IDENTIFYING COMMONALITIES

While we focus attention on proving programmatic success or failure, we have not put enough money or brainpower into looking across interventions, funders and sectors to find the commonalities of success. There are a few exceptions. In a study of 548 individual programs aimed at reducing recidivism among delinquent youth, the Center for Juvenile Justice Reform at Georgetown University showed it is possible to expand conceptions of how research should be used “without wavering in [the] commitment to evidence.”¹⁸ The Center’s researchers classified the interventions into categories, such as family therapy, cognitive-behavioral therapy, mentoring, social skills training, behavior management, individual counseling, and the like. Into these categories fell both nationally named programs and home-grown ones that had not been packaged for broader dissemination but were nonetheless recognizable examples of a common program type.

The Georgetown group examined research on the effectiveness of each type of program (measured by rates of recidivism) and found considerable variation in the program features, settings, personnel, and characteristics of the juveniles served. The researchers concluded that when evidence shows generally positive effects, that finding is in many ways more robust than the findings of the few studies supporting a particular model program.

On the other hand, the broad range of programs within each type means there is more variability in the research findings on those programs. For example, some family therapy programs show much larger

effects in some circumstances than others; indeed, some studies showed no effects or even slightly negative effects. Some home-grown programs are more effective than the name brands, and vice versa.

To use such evidence to guide practice, one must know which characteristics distinguish the more effective programs. So the researchers looked for a systematic pattern of factors associated with effectiveness that would extend “evidence-based practice” beyond the brand-name models to the locally developed programs whose effectiveness also is supported by research. Indeed, much of the programs’ effectiveness could be accounted for by a relatively small number of quite straightforward factors—like the decision to target high-risk cases and to take a therapeutic approach to changing behavior rather than a control or deterrence philosophy. This meant that close attention to these factors in the selection and implementation of programs for juvenile offenders could provide reasonable assurance that those programs would help reduce recidivism.

The Georgetown meta-analysis underscores the fact that while each individual program study is valuable in its own right, it becomes more valuable for the field when its findings are synthesized with findings from other studies to support broadly applicable, evidence-based recommendations. Child Trends has made an outstanding contribution to this category of research with its Fact Sheets, which synthesize the essential attributes of effective interventions on topics ranging from home visiting to adolescent reproductive health.¹⁹ Child Trends President Carol Emig says this kind of cross-program synthesis is a good way to improve programs: “Instead of telling a city or foundation official that they have to defund their current grantees because they are not evidence-based, funders can tell long-standing grantees that future funding will be tied at least in part to retooling existing programs and services so that they have more of the elements of successful programs.”²⁰

This is also why Grantmakers for Effective Organization (GEO) encourages foundations to gain insights about the commonalities of various solutions, not only among their own grantees but also other funders.²¹

It is tempting to think about common factors mainly in programmatic terms, but effectiveness depends as much on the quality of implementation as it does on programmatic content. Given this fact, there has been surprisingly little systematic effort to explore across domains what makes for successful implementation. Kristin Moore, the former president of Child Trends, led one such exploration. Her team examined experimental, quasi-experimental, and non-experimental research in addition to provider wisdom. In their report, “Program Implementation: What Do We Know?” the researchers identified implementation features that enhance program effectiveness as well as program features that do not work and a few that appear to be harmful.²² As more evidence about *what* has worked, and *how* is synthesized and used, both implementation and scale-up will be strengthened.

IMPROVING REPLICATION AND SCALE-UP

The history of replication and scale-up efforts is discouraging. As President Bill Clinton frequently declared, “You can find virtually every problem in our country solved by somebody somewhere in an astonishingly effective fashion....The challenge for us is to figure out how to replicate that.”²³ The day before his first inauguration, Clinton told the nation’s governors that his number one disappointment as governor had been that it was so hard to “take something that works to the next level.” He said he “could never figure out a way to make the exception the rule, and that is our enduring problem in America in public life.”²⁴

President Obama also sees replication as a challenge but seems to have more confidence that it can be done, even when it comes to an initiative as complex and multi-faceted as the Harlem Children’s Zone. During the campaign, he promised, “When I’m President, the first part of my plan to combat urban

poverty will be to replicate the Harlem Children's Zone in 20 cities across the country.”²⁵ That is the initiative we now know as *Promise Neighborhoods*.

The ascendant belief in the 1960s and 1970s was that successful programs carried the seeds of their own replication. Foundations hoped that when they funded a pilot that worked, it would be picked up by others and supported with public funds and reformed policies. By the end of the 1980s, it became clear that successful programs were dying when the demonstration funds ran out at much the same rate as those that didn't work.

Furthermore, when successful programs have been implemented away from their places of origin and taken to scale, they have at best modest results.^J As Dr. Jack Shonkoff has pointed out, “demonstration projects tell us what is possible, but their replication on a broader scale has been remarkably uneven.”²⁶ The desirable effects on delinquency and subsequent offending identified in research studies have often been found to be attenuated when those programs are scaled up for general application.²⁷ Another example comes from a study of the scale-up of the successful “systems of care” model, in which Jane Knitzer and Janice Cooper found that even with consistent adherence to “systems of care” principles and on-the-job training of practitioners, “systematic implementation was not assured.” Most discouragingly, there were no improvements in children's outcomes.²⁸

We have reason to believe that scale-up efforts^K have a brighter future, however, now that considerable experience and analytic power has been devoted to trying to understand and overcome the weaknesses of past efforts.²⁹ (See Fig. 3 for a description of identified flaws.) Moreover, the work that has gone into identifying reasons for the disappointing results of past scale-up efforts is now increasingly accompanied by thoughtful analyses of how to overcome past weaknesses, and actions that have succeeded in doing so.

^K We refer here to scaling up rather than replication because it is a more expansive term. Although the two words are often used interchangeably, replication is more likely to refer to a number of iterations of a single model, while scale-up may include replication, but also expanding impact by reaching larger numbers in other ways, such as by adapting a model or spreading an idea, a vision, or a framework.

Fig. 3: FLAWS IN PREVIOUS REPLICATION AND SCALE-UP EFFORTS

Research and experience suggest that the most frequent problems in scale-up efforts are:

Insufficient understanding of what makes the original intervention successful, of what must be held constant and what can be adapted. The alternative to providing this information to the replicators, has been requiring scale-up with fidelity to the proven model, using the originators' protocols or program manuals. Then the implementers are constrained from adapting the program to new circumstances.

Insufficient care and resources devoted to the quality of implementation in the process of scaling up. Psychology professor William R. Miller of the University of New Mexico and his colleagues identified challenges to the replication of model programs for substance abuse treatment. "With little understanding of the difficulties associated with quality implementation, practitioners who adopt these programs frequently find their efforts poorly supported," they found. "The clinician (or worse, the clinician's supervisor) may have attended a presentation on an evidence-based program at a conference or read a book written by the program developer. That experience is then followed by the clinician's attempt to use what was heard or read, perhaps as interpreted by a supervisor. This is not only a constricted view of how to go about evidence-based practice but an underestimation of the power of inertia in clinical practice."³⁰

Insufficient attention to the culture within the helping organization and the regulatory and systems context surrounding it.

Insufficient attention to local capacity and the organizational environment within which the intervention must be sustained.

Failure to recognize that what works for most children and families may not change outcomes for the children and families who are most at risk.

Failure to recognize the "uptake problem" among local front-line personnel and supervisors. These staff may resist abandoning their established practices to adopt a model program that was invented elsewhere. They may make their own adaptations to achieve a better fit with their real-life organizational constraints, such as large caseloads, little supervision, and resource limits on the types, frequency, and duration of services. If they have only a manual or protocol to guide them, they have little to go on to judge whether their adaptations will be more or less effective than the original.

Funders' reluctance to devote significant sums to the substantial operational costs of scaling up.

Until recently, there were basically two theories of how to scale up and replicate. The first emphasized the implementation of successful programs in new sites with full fidelity to the original. The second theory focused on duplicating the essence of a successful program while adapting many of its components to a new setting or population.

The rationale for implementing proven interventions in new sites with full fidelity to the original was laid out in 2009 by Public/Private Ventures (P/PV), an organization that has led many efforts to understand the triumphs and failures of replication efforts. "Program replication," wrote P/PV researchers, "is premised on the understanding that many social problems are common across diverse communities—and that it is far more cost-effective to systematically replicate an effective solution to these problems than to continually reinvent the wheel."³¹

The "pure" end of the scale-up spectrum, where the intervention to be replicated has been proven to work through random assignment experiments and will be replicated with detailed fidelity, is appealing because

the risk of failure seems minimal—especially when the alternative is posed as the arbitrary replication of some passing fad.

The rationale for replicating the essence of a successful program while adapting many of its components to new settings and populations is espoused by a number of experts, including Jeffrey L. Bradach, managing partner and co-founder of the Bridgespan Group. Bradach contends that the objective of replication is “to reproduce a successful program’s *results*, not to slavishly recreate every one of its features.”³² He argues that replication is anything but a cookie-cutter process; at its heart is the transfer of an organization’s theory of change to a new location. This might entail moving a few practices from one site to another or cloning the organization’s entire culture. Both approaches require considerable investment of funds, talent, and time,³³ and an array of evidence to help the organization determine how much fidelity or customization is needed to make the scaled-up model effective and how to differentiate the essential components from those that can and should be adapted to the new circumstances.

One excellent example of a successful approach to scaling up and spreading to other settings comes from the *Strive Partnership* of Cincinnati, which started with a comprehensive intervention in one city but allows for and supports adaptation as other cities implement the approach. *Strive* is a place-based intervention that pulls many diverse parts of the community together to pursue shared outcomes. It has united Greater Cincinnati leaders at all levels of the education, nonprofit, community, civic, and philanthropic sectors around shared issues, goals, measurements, and results. Seven outcomes drive the collective work and provide the framework for measuring success: kindergarten readiness, fourth-grade reading proficiency, eighth-grade math proficiency, high school graduation rates and ACT scores, and postsecondary enrollment and completion. These are the outcomes on which *Strive* measures its success. The guiding principles of the partnership include a focus on data driven decision-making, facilitating and sustaining coordinated action, and advocating for and aligning funding around what works.

Strive has developed a cutting-edge approach to scaling up. In response to requests to “replicate,” *Strive* is working with nine other communities to establish similar cradle-to-career initiatives and to achieve more rapidly the “collective impact” they are committed to. Rather than opening identical branches in other cities, *Strive* promulgates a flexible process for change, offering each community a common agenda, a short list of indicators for change at the community level and across participating organizations, and a set of tools adaptable to the community’s own needs and resources. With these tools and assistance from *Strive Cincinnati*, the new *Strive* communities take true ownership of their own initiatives but don’t need to start the process from scratch. Processes that took the original *Strive* several years to develop are being adapted and modified by other communities in significantly less time.³⁴

CHALLENGES TO REPLICATION AND SCALE-UP

A tension in choosing between ease of replication and potential impact on outcomes

runs through most of the advice provided by people who have studied the process of spreading success. If one program is easy to replicate because it is simple and circumscribed, and another is hard to replicate because it is complicated and hard to standardize, a market-oriented decision would surely choose the first for replication. But when we consider social purposes (e.g., What are we hoping to accomplish?), the question becomes more complicated. If the hard-to-replicate program is more likely to have a significant impact on outcomes, it may be preferable to struggle with all the complexities of trying to spread it.

If the hard-to-replicate program is more likely to have a significant impact on outcomes, it may be preferable to struggle with all the complexities of trying to spread it.

Similarly, if one successful program is easy to replicate because it operates outside mainstream systems, but another is hard to replicate because it partners with a public system, the social purpose may be better served by the second program's design. For example, when a foundation-sponsored community center found that its staff were called on to advise families involved with the child welfare system, the center sought to collaborate with the child welfare agency. Workers from both agencies ended up sharing office space and collaborating in providing services, which vastly complicated their intake protocols, their bookkeeping, reporting, and staffing practices. When they prepared to scale it up to other parts of the city, their arrangements with the child welfare agency made the replication plan more complicated but, they judged, better at serving the community.

Resources are another issue for efforts to scale up what works. An organization that attempts to scale up its own programs or someone else's has unique funding needs that rarely fit well with the inclinations of most public or private funders.

Spreading an effective program is a difficult and resource-intensive job. At a minimum, P/PV suggests, it requires implementers to: identify and document the essential and adaptive program elements; create implementation and training guides that encourage consistency across sites; establish dedicated staff positions to support the replication effort; develop a network of strategic partnerships that spans from the local to the national level; establish a universal data collection system to monitor results; and establish standardized training and ongoing technical assistance.³⁵

Bridgespan's Bradach believes that the failure to replicate innovative social programs often is "simply a problem of money." He points out that at the moment when large amounts of capital would flow to a proven idea in the for-profit sector, funders in the nonprofit sector frequently back away—victims of donor fatigue, a belief that equity necessitates spreading money around, hesitance to make 'big bets,'"³⁶ and a reluctance to support non-programmatic activities.

For the most part, the funding patterns of the nonprofit sector (i.e., small grants, for short durations, focused on discreet programs), can work against building organizational capacity. . By contrast, the grant-making practices that support scaling up involve providing general operating support; making multi-year commitments; and cultivating supportive, respectful relationships. The Edna McConnell Clark Foundation pioneered this approach to building organizational capacity for programs with success in improving outcomes and with a clear evidence base. Their work is influencing both public and private funders. The Clark Foundation and others, such as Venture Philanthropy Partners, strengthen successful organizations that in turn spread their programs, ideas, and practices.

Community capacity is a third dimension of spreading success that does not lend itself to the usual efforts to assess and replicate what works. The Aspen Roundtable on Community Change found that complex social interventions require significant capacity to implement—capacity that under-resourced organizations in distressed neighborhoods often do not have. An extensive review of two decades of community efforts conducted by the Roundtable found that "the lack of capacity in distressed communities makes it difficult to implement traditional programs effectively and virtually impossible to move from business-as-usual to something more ambitious."³⁷

Therefore, the Roundtable continued, deliberate investments must be made to build broad and deep capacity in which "ranging from ability to implement high-quality programs well, to ability to hold collaborators and systems accountable to the community, to ability to take advantage of opportunities and mitigate threats as they occur."³⁸ In fact, the lack of capacity has often turned out to be the most serious limitation on success in creating place-based and community-wide change.³⁹

Recent federal initiatives reflect a new awareness of the importance of community capacity as well. The federal guidance for *Promise Neighborhoods* and *Choice Neighborhoods* allow funding for data capacity, development of community collaboratives and other supports without which complex interventions are unlikely to succeed.

The replication of small-scale, circumscribed, proven programs, or even high-quality organizations, is unlikely to be the best single strategy for substantially improving outcomes for large populations of disadvantaged children and families. To improve outcomes at scale today we must look not only to individual breakthrough programs but to strategies that effectively target outcomes no single organization or program can achieve on its own.

To maximize outcomes for children and families, we need—along with identifying strong programs and replicating those that can stand alone—to focus more attention on identifying powerful practices that emerge from analyses across programs, which transfer more easily than do model programs. But we also need to use new knowledge from research and experience to increase the *effectiveness* of promising programs, by adding missing pieces, and combining and connecting them to each other and to systems that will support and sustain them. That will give us a chance to reach far larger numbers and, potentially, to improve outcomes for entire populations of children and families.

In its most recent report on this subject, P/PV maintains that confronting the problems we face in the second decade of the 21st century “calls not just for tested models and proven services.... [W]e 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.”⁴⁰

The unsolved “wicked problems” that face us today include racial disparities in health, high rates of academic failure and school dropout, teen births, asthma, substance abuse, obesity, family and neighborhood violence, child neglect, social isolation, and concentrated poverty. These problems are “not fully reducible to component parts”; they are caused by such complex forces that “tightly choreographed solutions will not change their course.”⁴¹

In the business sector these are known as “adaptive problems.” They can’t be solved by holding proven programs constant, or by isolated initiatives because no single entity has sufficient resources or authority to bring about the necessary change.⁴² These are problems that “do not have a singular causal agent, a straight-forward mechanism of causation, or a relatively consistent result for all who would receive a standardized intervention.”⁴³ As Gara LaMarche, president of Atlantic Philanthropies, warns, “[O]rganizations working for social or policy change should understand that no significant change was brought about by one organization working alone.”⁴⁴

Unlike “technical problems,” which an expert or a circumscribed intervention can fix, adaptive problems require solutions that involve what John Kania and Mark Kramer call “collective impact.”⁴⁵ These solutions “require people in the community to change their values, their behavior, or their attitudes, (and) ... to learn new ways of doing business.”⁴⁶ And because these solutions pose a heavy implementation challenge, a focus on achieving and documenting effectiveness at scale is essential.

To achieve transformative social change, we must look not just to individual breakthrough programs but to strategies that effectively target outcomes no single organization or program can achieve on its own.

The most effective approaches to these adaptive problems tend to be population-based, multi-level, and multi-faceted. Thus our expanded evidence base must support:

- The non-programmatic parts of initiatives that involve systems, policies, community norms, and the infrastructure that ensures coherence and continuing quality improvement.
- The initiatives that involve a complex web of individual behavior, social norms, institutional supports, reformed policies, and wholesale changes in the physical and social environment
- The initiatives that require participation from actors in multiple sectors and are influenced by a constellation of complex social and political forces, some of which change during the process of solving the problem.⁴⁷

By increasing the effectiveness of current programs, adding missing components, linking them to each other and to supportive systems, and providing infrastructure to monitor, improve, and sustain them, we have the best chance of achieving transformative outcomes at greater scale.

Fig. 4 offers suggestions for how one federal program, the *Social Innovation Fund*, might use an inclusive approach to evidence to strategically scale up successful interventions.

Fig. 4:
**POTENTIAL IMPLICATIONS FOR SCALE-UP
RELATED TO THE *SOCIAL INNOVATION FUND***

The *Social Innovation Fund* (SIF) provides unprecedented opportunities for trying out and learning from improved scale-up strategies. This federal funding stream, enacted in 2009, seeks to: (a) recognize and increase the impact of social entrepreneurs and other nonprofit community organizations in tackling national and local challenges; (b) increase private and public investment in nonprofit community organizations that effectively address national and local challenges, to allow them to replicate and expand proven initiatives or support new ones; and (c) strengthen infrastructure to identify, invest in, replicate, and expand initiatives with effective solutions to national and local challenges. Strategically applying the information that comes from knowing what successful scale-up efforts have in common, as well as the characteristics of failed scale-up efforts, could significantly strengthen the implementation of SIF.

SIF's implementation strategy has put a priority on obtaining the most precise information possible from experimental evaluations so that funding can be more accurately focused on the best solutions. It is using the rigor of evaluations as a prime criterion for selecting the organizations (intermediaries) through which new federal and philanthropic money would flow, for selecting interventions (subgrantees) for replication and scaling, and to learn from the new activities funded by SIF.

We believe that a more inclusive definition of what counts as credible evidence would enrich SIF's strategy by allowing intermediaries and subgrantees to draw, build, and act on an expanded knowledge base. For example, SIF grantees could:

- Test a variety of criteria for judging which interventions are worthy of scaling up.
- Test the practicality and impact of replicating successful models with full fidelity to the original, compared with focusing on cross-cutting factors that make interventions successful; add to the understanding of which circumstances of interventions are most suited to “pure” replication and which are likely to require more adaptation.
- Test a variety of approaches to implementation and scale-up that would improve outcomes among the children and families most at risk.
- Generate new evidence about how to maintain high quality in implementation and scale-up and how to reconcile the pressure to innovate with pressures to implement only those programs and practices that are already proven.
- Generate new evidence about how best to gauge accomplishments and document them.
- Document the extent to which SIF grantees and subgrantees do indeed bring a new, entrepreneurial approach to national and local challenges (i.e., greater flexibility, fewer bureaucratic obstacles, new kinds of problem solving).

SECTION 3: ADOPT A PRAGMATIC APPROACH TO ASSESSING COMPLEX INTERVENTIONS

In the past, well-established and trusted methods of learning and evaluation focused on **interventions** operating in simpler contexts than those addressing today’s complex, interrelated social issues. As we look toward the next generation of social interventions, we need solid, credible, and deep evidence about what works to solve complex social problems. This means devising new and better ways of locating, assembling, and evaluating evidence from the workings of interrelated interventions and strategies.

Yet, too often, current pressures run in the opposite direction, setting up tensions between funders, researchers, and practitioners. We believe those tensions can and must be resolved. To that end, in this section we explore why practitioners, managers, and funders are struggling with more, rather than fewer, constraints on what is considered credible evidence; examine the limitations of traditional evaluation methodologies when applied to more complex initiatives; endorse a framework that focuses on learning in a variety of ways; and outline a pragmatic approach to evaluating complex efforts. The next section of the paper (Section 4) follows with recommended methods for generating “real-time,” results-based learning through close tracking of results as part of complex initiatives’ accountability strategies. Information generated in that way, when accumulated and analyzed, can help close the gap between current evaluation strategies and the information needed to improve outcomes.

As we look toward the next generation of social solutions, we need solid, credible, and deep evidence about what works to solve complex social problems. This means devising new and better ways of locating, assembling, and evaluating evidence from the workings of complex interventions and strategies.

CONSTRAINTS ON WHAT IS CONSIDERED CREDIBLE EVIDENCE

The Obama Administration signaled its intent to insist on clear demonstration of past effectiveness as a prime criterion in allocating federal resources in June 2009, when Peter Orszag, director of the Office of Management and Budget, declared that “to be considered for funding, [organizations] must provide credible evaluation results.”⁴⁸ In laws, regulations, and guidance promulgated during the year that followed, the guidance about what were to be considered credible results focused more narrowly. Most of the directives that fleshed out the rules for funding new initiatives prioritized evidence from randomized trials and other experimental methods as preconditions of eligibility.

This direction was widely applauded, especially in Washington, DC. As faith in government and other large social institutions diminished, as resources for social programs became increasingly scarce, concern spread that policymakers might be conned into supporting programs that would ultimately fail. Stories circulated about promises of effectiveness made by researchers who were sloppy with their data, wore rose-colored glasses, cherry-picked the data, or had become advocates. A clear focus on supporting only what was known with certainty to have worked in the past seemed a welcome way of minimizing risk.

Orszag explained to sympathetic listeners in all three branches of government that he was committed not only to funding what has been shown to work but also to stopping funds for what doesn’t work, and that the only protection against funding what doesn’t work is experimental proof.⁴⁹ Unless the effectiveness of an intervention can be quantified, and the cause-and-effect relationship between the intervention and

the apparent results demonstrated by use of a counterfactual, the reasoning went, it would be unscientific—and therefore irresponsible—to continue or expand its operation.⁵⁰

This argument reflected the prevailing sense that we could evaluate social programs with the same methods that had led to the nation's great medical advances, that the best information came from studies conducted under laboratory conditions, and that effective programs could show that a “measurable amount of social benefit” was attributable to a particular intervention.⁵¹ A Brookings Institution forum concluded that even when laboratory conditions cannot be maintained, what works in educational and social programs can be identified with certainty through randomized *field* trials in which “one person gets the pill, and the other person gets the placebo.”⁵²

The idea that you could actually pick social interventions that had been *proven* to work using scientific methods was not only reassuring but thrilling. Funders, whether public or philanthropic, didn’t have to make fallible judgments—they could rely on proof that came out of incontrovertible numbers. As economist Rob Hollister says, randomized clinical trials are “like the nectar of the gods: once you've had a taste of the pure stuff it is hard to settle for the flawed alternatives.”⁵³

LIMITATIONS OF CURRENT EVALUATION TECHNIQUES

Traditional study designs are generally linear; they tend to reduce a problem to a single or a smaller number of causes in order to apply experimental design methods.^L Yet, in one field after another, people have grown aware that so much of what seemed to work in social programs and social policy was complex, multifaceted, and evolving.^M We are learning from a wealth of experience that improving results, especially for the most vulnerable, means changing many things at a time.

An example comes from the Harlem Children’s Zone (HCZ), working to reweave the social fabric of Harlem that was torn apart by crime, drugs, and decades of poverty. HCZ has developed programs to address health, education, early childhood, family support, and community building, all of which continually evolve in response to data that demonstrate where outcomes are being achieved and where improvements or additions are needed. HCZ’s infrastructure, which accounts for the whole being more than the sum of its parts, seeks to ensure that all the children in the Zone stay on track from birth to college and entry into the job market and that when new problems are discovered, the many-faceted responses that are required can be mobilized.^N It is the impact of these interactions and this infrastructure that is so difficult to evaluate by traditional methods.

^L This process is referred to as “reductionist research,” defined as based on a “model that approaches statements of causality by isolating, simplifying, and holding constant key conditions as an attempt is made to understand effects by controlling or removing all potential confounders.” Kumanyika and Institute of Medicine, *Bridging the Evidence Gap in Obesity Prevention*.

^M “Reforming public education ...and improving community health are all adaptive problems. In these cases, reaching an effective solution requires learning by the stakeholders involved in the problem, who must then change their own behavior in order to create a solution.” Kania, J. and Kramer, M. (Winter 2011). “Collective Impact.” In [Stanford Social Innovation Review](#).

^N For example, when a wide-ranging household survey conducted by the Harlem Children’s Zone found that a third of the neighborhood children under age 13 tested had asthma—more than five times the national average—HCZ enlisted a wide range of partners in its efforts, which drastically decreased hospitalizations, emergency room visits, and school absences for children in the program. Harlem Hospital provided the medical knowledge and medical care, including trained health staff that made home visits every three months. The Columbia University School of Public Health supported data collection and evaluation. New York City’s Department of Health and the Urban Planning Program of Columbia’s Graduate School of Architecture, Planning, and Preservation offered technical assistance on the environmental aspects of the program. New York Legal Services offered legal assistance. And HCZ, with its deep ties to the community, reached out to neighborhood families, who trusted HCZ enough to assume the responsibilities that families have to help reduce the incidence and severity of their children’s asthma.

For some time, randomized trials and other experimental evaluation methods have been a poor fit with complex interventions, viewed as too expensive, too time-consuming, and failing to offer much of the needed information. When the problem does not have a single cause or single solution, these evaluation designs fail to shed light on enough of the problem and its potential solutions to provide the guidance that is necessary for decisions about the overall intervention. As Feather O'Connor Houstoun, president of the William Penn Foundation, has observed:

Ironically, the rigor of the control group trial design may be its Achilles' heel. The more a single intervention is isolated from other elements of a program for purposes of testing its singular effectiveness, the more its interaction with other program features remains obscured and unexamined as part of the program evaluation.⁵⁴

Under increasing pressure by government and philanthropy to obtain proof of what works from experimental studies of social programs and policies, however, some academics, advocates and practitioners did question whether the great contribution made during the 20th century—testing medical procedures and products with experimental methods—was likely to be appropriate for testing complex social interventions and policies in the 21st century. Social and educational reformers became more open about their misgivings when even distinguished leaders in medicine began to question the supremacy of the Randomized Clinical Trial (RCT) for understanding a full range of interventions.

Donald M. Berwick, the crusading physician who heads the Centers for Medicare and Medicaid Services, wrote that the RCT is a powerful, perhaps unequaled, research design, but only to explore the efficacy of selected *components* of practice—drugs, procedures, and other interventions that are conceptually neat and have a linear, tightly coupled causal relationship to the outcome of interest.⁵⁵

In a 2010 report, the Institute of Medicine (IOM) working group assembled by the National Academy of Sciences to inform decisions about obesity prevention called on the scientific community “to rise to the challenge and transform the evidence picture to be commensurate with the needs of policy makers and funders.” The IOM group concluded that a more useful approach would expand evidence by crossing disciplines, making connections among population-level and community-level influences, obtaining the practice-based evidence that is intrinsically relevant to natural settings, “focusing on the whole picture and not just a single element,” and “thinking about...community-level influences.”⁵⁶

Similar questions about the over-use of experimental evaluation methods have been expressed in the domains of early childhood, education, child abuse and neglect, and international development. The Government Accountability Office (GAO) has warned against OMB's elevation of experimental evidence to the top of the methodological hierarchy. In a report released in November 2009, the GAO concluded that requiring evidence from randomized studies as the sole proof of effectiveness will likely exclude many potentially effective and worthwhile practices.⁵⁷

Perhaps even more persuasively, the American Evaluation Association, commenting on the approach used by the Department of Health and Human Services to select “evidence-based” models, recommended that

“The RCT is a powerful, perhaps unequaled, research design to explore the efficacy of conceptually neat components of clinical practice—tests, drugs, and procedures....[But] we have overshot the mark. We have transformed the commitment to ‘evidence-based medicine’ of a particular sort into an intellectual hegemony that can cost us dearly if we do not take stock and modify it.”

—Berwick, D.M. (2005). “Broadening the view of evidence-based medicine.” *Qual. Saf. Health Care* 14:315-316.

the selection process: “(1) forego assigning an automatic high rating for random assignment designs and automatically relegating all other evaluation designs to moderate or low ratings, (2) avoid using the label ‘gold standard’ in connection with random assignment designs in the rating methodology, (3) use additional criteria to assess the value of impact evaluations, (4) more specifically identify alternative impact evaluation methods, and (5) emphasize the value of multiple studies and mixed methods.”

In this atmosphere, front-line workers, managers, and even some funders began to be more vocal about problems with using the experimental methods seen as “most rigorous” to determine the success and understand the workings of social and education programs and policies. They began looking for alternatives that would enable them to:

- Adapt their interventions to “refine strategy and improve implementation over time,”⁵⁸ reflecting advances in knowledge, changes in context, evaluation findings, and experience, thus freeing them from the experimental necessity of holding the intervention constant;
- Avoid the large costs of randomized trials;
- Obtain measurable results promptly enough to be relevant;
- Learn not just whether but *how* their work affects outcomes;
- Learn about interventions that are complex, interactive, and relationship-based; that can be adapted to a variety of cultures and populations as well as new and changing contexts; and require significant front-line flexibility and sensitivity;
- Learn about reforms needed at the institutional, policy, systems, and population level;
- Learn about the impact of reforms that do not consist of isolated programs operating within isolated silos but combine proven and promising practices in new ways; and
- Learn about interventions too young or small to be assessed with experimental methods.

The key to overcoming these limitations is not to give up trying to assess impact but to stop insisting on *proving* the impact of every kind of intervention. When all participants develop more capacity to articulate the logical links between strategies used and the results measured, we will be better able to document and understand impact. This kind of learning will support better decision making, implementation, testing of theories of change, and production of knowledge for the field.

We recognize that alternative approaches to evaluation would balance the advantages that come with certainty about one class of interventions (those with a direct and linear causal relationship to the outcome) with the advantage of gaining more usable knowledge about interventions that are more complex but carry a greater probability of improving outcomes.

The key is that neither policy nor practice should be driven by evaluation methods. Instead, evaluation should be driven by policy and practice needs. Research methods should match the policy and practice questions to which we need answers, and the nature of the interventions about which we need to learn about.

The key to overcoming these limitations is not to give up trying to assess impact but to stop insisting on proving the impact of every kind of intervention.

A PRAGMATIC APPROACH TO EVALUATING COMPLEX EFFORTS

The approach to evidence and evaluation that we promote is informed by accumulated evidence, recent experience with complex initiatives, and the insights of expert observers such as Grantmakers for Effective Organizations, the Aspen Institute’s Roundtable on Community Change, and the Institute of Medicine’s obesity prevention study group.

This approach starts from several premises. First, it assumes that it is possible to assemble data and information, even about complex interventions, that enable policy makers to make informed judgments, based on strong evidence, about which interventions are most likely to be effective and which are probably less effective.

Second, the approach posits that a broader collection of evidence is likely to stimulate effective action rather than cynicism about what can or can’t be done. As Dr. Donald Berwick suggests, “People committed to science ought to supply not just skepticism but also hope. People committed to science ought not just to judge change but to lead change, not just to evaluate the rest of us but to join the rest of us.”⁵⁹

And third, this approach recognizes that evaluations of complex interventions usually have to serve multiple purposes and provide guidance on a range of issues. Policymakers seek answers to multiple questions about complex interventions, and evaluators must be able to contribute useful information about many if not all of them. As Tom Kelly of the Annie E. Casey Foundation says, “Policymakers are interested not just in the question, ‘Does this intervention work *in this place, in this context, and with this population?*’ but are also trying to answer the questions around *what* and *how* similar successful interventions *can be adapted, replicated and expanded* in other places. This requires the evaluation to generate useful and timely information about implementation, about process, and about the interrelationships among multiple parts of an initiative. In short, what policymakers and funders want to know is not simply, ‘*Did the initiative work?*’ but also, ‘*What can you tell me about what worked here that can be useful in applying a similar approach there?*’ ”

Given these multiple demands, we suggest an approach to evaluation that begins with the ends in mind; is as clear as possible about the theories of intervention; uses multiple methods of evaluation, adapted to the purpose and nature of the intervention; and uses experimental methods when they are appropriate and in the context of a broader approach to knowledge development. Six elements of this approach are particularly important.

1. Begin with a results-oriented framework.

Identifying the clear, measurable results sought by a complex intervention is the essential first step toward both successful implementation and to a successful evaluation. The use of a results framework from the beginning provides clarity of purpose and also helps build a commitment to data, accountability, and performance. By the terms *results* or *outcomes*, we mean improvements in the well-being of children, families or communities—the real ends of complex initiatives—rather than an initiative’s influence on practice, systems change or even policy change, vital as those are. A focus on measurable well-being inspires multiple stakeholders to act together (essential for the improvements in whole populations sought by these initiatives), supports the case for multi-year investment, and keeps people attending to what it will take to achieve ambitious, long-term goals. As Fiester and Smith observed about the Casey Foundation’s *Making Connections*:⁶⁰

The results focus can be a formidable and sustainable force for change. [Results] serve both as the driver of and glue for our work within sites....In many respects, it is the commitment to results that brings diverse people together, supports collaboration across boundaries, instills a

shared sense of ownership for what happens in the community, and mobilizes specific actions. It is the achievement of early results that gives people confidence that things can move in a positive direction.

A results framework is of heightened importance in complex interventions because these require collective action by multiple partners in order to succeed. Such action is difficult to inspire and seems impossible to maintain unless it is rooted in a shared commitment to results that make the hard work of joint action worthwhile.

There are many challenges to developing and sustaining the type of “results culture” that allows multiple stakeholders to work together over many years. Among them are the difficulty in achieving or observing changes in child and family well-being in a short period of time; the tendency to set targets for change that are either too ambitious or not ambitious enough, thus undermining attention to results measurement; the inadequacy of systems and tools for measuring and tracking results; and the difficulties of maintaining commitment to continuous improvement across multiple organizations and systems.⁶¹ However, the demand for results-oriented approaches is stimulating more results-oriented research, and more attention to the need for work involving both practitioners and researchers in developing more useful ways of measuring results (to which we return in a later section).

2. Use strong theory to connect activities to results.

A results frame that is based on strong theory and draws on research and practice to connect interventions to results turns evaluation into a practical tool for measuring the impact of complex social interventions. Theory-based evaluation “acknowledges the importance of substantive theory, quantitative assessment, and causal modeling but does not require experimental or even quasi-experimental design.”⁶²

In theory-based evaluations, the implementers of a complex intervention identify the components that they intend to use or strengthen to contribute to the agreed-upon results for children, families and communities by “mapping backwards” from target results to the specific practices, programs, strategies, opportunities and resources that are expected to produce the results. This process produces a theory of change which identifies the components that must be present for the expected outcomes to occur. It also specifies the various characteristics of the intervention, target population, and context. A well-crafted theory of change also helps focus data collection on what matters most, thus saving the initiative from a wasteful collection of information about everything. As Atlantic Philanthropies President Gara LaMarche reminds us, “Evaluation should measure only what is important.”⁶³

In complex interventions, theories are strongest when they are specific (and research and experience-based) about program components, processes, interrelationships, and dosage, reach and saturation. Complex interventions are typically aiming for impact on hundreds and eventually thousands of children or families, not just proving Program A works for 50 children, for example. This requires specificity about a theory of scale and a theory of sustaining impact—that is, about how a community can continue to achieve the results it seeks. Theory of this depth requires unusually close work between implementers and evaluators, and is likely to evolve considerably over the life of an initiative.

Theory, perhaps paradoxically, also helps evaluators take into account the complexity of real-world contexts in which interventions are implemented and decisions are made. Theory and experience

combine to guide evaluators to focus “on the whole picture ... the wider context ... interactions among different components [including] community-level influences ... and practice-based evidence.”⁶⁴

Public health successes, especially in reducing tobacco use in the United States, confirm the usefulness of this broader perspective. Tobacco control efforts initially targeted individuals and their behaviors, but evolved to a multi-level systems approach. Individual behavior change was the goal, but the strategies that worked involved industry, legislation, public health programming and messaging, and the health care system. State by state, it became clear that more complex combinations of strategies were better than any single intervention, and the more components devoted to the clearly defined result the better. The two states, California and Massachusetts, that undertook the most comprehensive interventions doubled and then tripled their annual rate of decline in tobacco consumption relative to the other 48 states. The other states showed more interest in the data from these two states than from several thousand controlled trials focused on individual behavior change in the scientific literature, because these two states had actually been able to achieve change at scale.⁶⁵

3. Expect to compare results, but don't expect to find a perfect comparison group to “prove” causality.

It is essential to have some way of comparing results to establish—not with certainty but beyond a reasonable doubt—that the observed change has a high probability of resulting from the practices, strategies, and programs under consideration (and not from factors, such as selection bias, that produce markedly non-comparable populations). The community-specific nature of place-based interventions makes it very hard, if not impossible, to find a comparison group that would allow for precise causal attribution. Furthermore, many change initiatives operate in neighborhoods with highly mobile populations, and the study sample is unlikely to stay in the neighborhood over the life of the evaluation. And, because many place-based interventions attempt to reach all members of a community, or all high-risk subpopulations, it is not feasible to randomly assign individuals to treatment and control groups. As the Annie E. Casey Foundation’s Tony Cipollone puts it, “Trying to do something as sophisticated as random sampling, and crafting an evaluation so airtight that it establishes causality—that's fine for evaluating a programmatic intervention but not for community change work.”⁶⁶

One option is to use empirical outcome data that show the impact of interventions for populations being served on widely sought, publicly reported results such as school readiness, absence of preventable health problems, academic achievement, and graduation rates. Using these data, it is possible to compare outcomes among the populations of children and families served by a specific initiative to (a) similar populations in the geographic area before the intervention began for whom baseline data are available, (b) current populations that did not receive similar services and supports, but for whom data are already available, i.e. does not have to be collected as part of the evaluation, and (c) national, state or local norms. These kinds of comparison groups constitute a “counterfactual” that does not offer mathematical certainty but does provide extremely useful knowledge when applied with intelligence and common sense.

4. Use multiple evaluation methods that align with the multiple purposes of evaluation, the nature of the intervention, and the stages of implementation.

For complex social interventions, multiple methodologies are necessary to obtain the range of learning, information, and evidence that they can generate. The complexity of these interventions and the range of possibilities for knowledge development they present require clear-eyed decisions upfront about the most important purposes to be served by evaluation, the most important questions to be addressed, the best match of methodologies to the nature of the intervention, and at what point in the initiative various interventions will be evaluable.

Among the typical goals of evaluation are to: (1) allow stakeholders (including funders) to judge whether the practice, program, initiative, or strategy is making progress toward the stipulated results; (2) allow implementers to decide how to improve implementation and performance in order to improve outcomes; and (3) inform the field about lessons learned and insights gained from the experience of implementing the initiative. In complex social interventions, evaluations are expected to achieve all three of these purposes (in some proportion), and the art of the evaluator is to determine which combination of methodologies can best meet these interlocking expectations, given the initiative's nature and its stakeholders' expectations.

Thus, in assessing whether the practice, program, initiative, or strategy is achieving results for children and families or making progress in that direction, a purely descriptive or formative evaluation will not be sufficient, although it is essential in an initiative's early stages. If evaluation findings are to be used for decisions about resource allocation, then it is important to document improved results for children and families, even if the results' connection to the intervention cannot be proven conclusively and even if the ultimate outcome will not be attained in the short run. The focus in such an instance must at least be on collecting interim measures that are known to be related to later results (e.g., fewer multiple placements of children in foster care predict better adult outcomes for children in the child welfare system).

For the purpose of improving implementation, quality, and performance in order to improve outcomes, evaluation methodologies must ensure that crucial elements of implementation are identified and how they connect to the expected outcomes is specified. If the intervention is using a variety of approaches within a single organization or among several organizations that share the same goal, the evaluation should shed light on the comparative effectiveness of each strategy—and should produce this information in “real time” so that program improvements or adjustments can be made. In these instances, qualitative methodologies will be important, so that the decisions and action steps by which interventions are adapted for differing conditions, populations, and problems can be teased out, generalized, and translated into advice for future implementers.

The multiple methodologies suggested above will be necessary to achieve an evaluation's purpose of informing the field and generating lessons learned from the cumulative experience of an initiative's implementation. In addition, evaluators will need to identify the characteristics of the initiative's process and outcomes that can be generalized to different circumstances.

The appropriateness of different methodologies will also evolve according to the stages of a complex intervention. Formative and descriptive evaluations are useful early on, as initiatives are shaped and when stakeholders are coming together, agreeing on ideas (and refining desired results), and putting in place the skills and capacities required to achieve the desired results. Beyond these initial stages, evaluation methodologies must assess whether capacities were actually put into place and document early program accomplishments. Timely feedback from evaluation during this stage is particularly important: accurate assessment of whether the prerequisites for larger-scale change are in place can head off too-early expansion of strategies that have not yet demonstrated success.

Only as implementation becomes robust—usually after three to five years—are impact evaluations appropriate or productive. And finally, as complex initiatives move to greater scale, evaluators must assess the extent to which the outcomes at scale match those achieved in pilot efforts.

Choosing the combination of methodologies that will achieve these multiple goals, for differing types of interventions and at multiple stages, constitutes the challenge in evaluating complex interventions. Our point in arguing the case for multiple and flexible methodologies is to challenge evaluators and implementers, together, to spend the time at the beginning of an initiative to

make choices about *what it is most important to learn*, and then to choose the evaluation *methodologies that provide the greatest likelihood of generating that learning*. In that spirit, among the methods that evaluators and implementers are finding most helpful to assess complex place-based efforts are:

- Qualitative studies and process evaluations, documenting the way in which ideas, resources and stakeholders are brought together to launch and implement an initiative;
- Cohort studies, to look more intensively at the results for specific target groups of children or families within the broader population whose needs are addressed by a complex intervention;
- Neighborhood surveys of impact, to begin to get data on population-level results; and
- Case studies of specific interventions, to understand in depth the way in which they do or do not affect change.

It is the skillful and insightful combination of these methodologies around a core set of learning questions and within a culture of “learning for accountability” that determines whether evaluation contributes the fullest range of knowledge possible.

5. Use randomized experimental evaluation designs when the focus is on determining with certainty whether or not an intervention that is clearly defined with a specific mode of action and effect has achieved the stipulated result.

Randomized experiments can be powerful methodologies in establishing proof of effectiveness in defined circumstances. As the Government Accountability Office has concluded, they are “best suited for interventions in which exposure to the intervention can be controlled and the treatment and control group’s experiences remain separate, intact, and distinct throughout the study.”⁶⁷ This is why we use randomized trials to determine the safety and efficacy of drugs. This is also why it is so difficult to evaluate complex interventions like the Harlem Children’s Zone (HCZ) with experimental designs. In fact, HCZ points out that none of its programs, except for its schools,⁰ have been subject to randomized experiments because “such a research design and the attendant denial of services (whether temporary or permanent) associated with it are inconsistent with HCZ’s mandate to serve all of the imperiled children living in our designated area.”⁶⁸

Within the context of a complex intervention, controlled experiments using random assignment of participants may be possible for one or more individual program components that are clearly and fully specified, that are held constant over time, for which the key variables can be defined and measured, and where the program is closely linked in theory and in practice to the target outcome. In considering this option, however, the question would be whether the cost and implementation challenges of an experimental approach would be justified by the additional knowledge generated, and whether the program component so studied is one that would remain relatively consistent as the overall intervention is scaled up or adapted to other locations.

⁰ An experimental evaluation of HCZ’s Promise Academy Charter Middle School found that by the end of eighth grade (three years after random assignment), lottery winners (students admitted to the school) were 27 percent more likely than lottery losers to be on grade level in math, and lottery winners were 15 percent more likely than lottery losers to be on grade level in language arts.

6. *Use non-experimental evaluation designs when causal connections are diffuse and when the prime focus is on understanding how, why, and the extent to which multiple interventions contribute to the stipulated results .*

As the Aspen Roundtable puts it, “We need new ways of learning when causal connections are diffuse and difficult to establish.”⁶⁹ This may occur in:

- Complex programmatic interventions that are interactive, and operate at several levels and in diverse domains,
- Interventions that need to adapt and evolve in response to new and changing contexts, lessons learned, and advances in knowledge, and
- Complex initiatives that involve systems and policy change, build infrastructure to ensure coherence and quality, and aim for change across an entire neighborhood.

In each of these circumstances, it is the complex, interrelated, and evolutionary nature of the interventions that make them so difficult to evaluate using experimental methods. This may be why, although these are precisely the kinds of interventions most likely to bring about the social changes now needed, the social sector remains focused on the isolated interventions of individual organizations. As John Kania and Mark Kramer observe, many funders try to ascertain which of many grant applicants make the greatest contribution toward solving a social problem, while grantees “compete to be chosen by emphasizing how their individual activities produce the greatest effect”:

Each organization is judged on its own potential to achieve impact, independent of the numerous other organizations that may also influence the issue. And when a grantee is asked to evaluate the impact of its work, every attempt is made to isolate that grantee’s individual influence from all other variables.⁷⁰

Ultimately, Kania and Kramer found, “funders and nonprofits alike overlook the potential for collective impact because they are used to focusing on independent action as the primary vehicle for social change.” This mindset was illustrated at a 2009 meeting at the Brookings Institution to discuss the impending Obama Administration initiative to fund 20 or so communities to undertake comprehensive place-based initiatives inspired by the Harlem Children’s Zone. Each speaker emphasized the important role of evaluation in holding the recipients of federal funding accountable and ensuring that the field would learn from the sites’ experiences. But when someone asked how the evaluation could be done, a panelist replied without hesitation that each of the programs that made up each community effort would be evaluated to determine how effective it was. No one in the room seemed disturbed that this evaluation approach would miss the most promising and unique aspects of the proposed initiative, which had to do with the impact of many related programs working together, supported by an infrastructure and policies that aimed to change an entire neighborhood and the context within which programs operated.

This perspective—focusing primarily on specific components of complex interventions—is in stark contrast to one of the findings of *Voices from the Field Vol. III*. Patricia Auspos explains that the community change efforts that progress most smoothly have the capacity to judge what is missing and how gaps can be filled, to provide essential data and analysis infrastructure, to connect the local efforts to regional and national resources, and to wrest a coherent whole from individual programs.⁷¹

If, as Kania and Kramer suggest, “It is no longer enough to fund an innovative solution created by a single nonprofit or to build that organization’s capacity,” we need a mindset change. We have to stop assuming (as many current federal requirements do) that just because an initiative can’t be assessed experimentally today, it surely can be tomorrow. That is only true of efforts that are too small or too new

for experimental evaluations. It is not true of those that cannot be evaluated experimentally because they are too complex and require collective action to achieve results.

We also need a change in the evaluators' roles and in the tools available. Too few tools exist to measure the impact of complex interventions, including place-based change efforts, multiple changes caused by multiple interventions, and the impact of systemic change that “encompasses the whole picture, highlights the broader context, considers interactions among multiple levels, recognizes the dynamic shifts that occur over time, and encourages collaboration among investigators from a variety of disciplines.”⁷²

In sum, an effective approach to evaluating complex interventions—both the programmatic and non-programmatic components—requires multiple techniques and multiple ways of learning. Currently, the strength of our conviction about the need for new combinations of evaluation methodologies and ways of learning exceeds our ability to describe in full detail what these look like. One promising approach emerging from complex place-based initiatives, however, is to incorporate systematic learning into the management of these efforts by a “feedback” loop generated from regular and thoughtful tracking of how the combined interventions are (or are not) producing the desired results and the requisite community capacities.

In Section 4, we turn to those management and accountability practices that can contribute to learning.

SECTION 4: CREATE AN EXPANDED LEARNING FRAMEWORK AND MANAGE TO RESULTS

The idea that nothing is worth knowing unless you know it for certain has its place, but not when applied to complex social interventions and policies.^P We can learn so much, including about program effectiveness, if we don't insist on absolute proof. Several influential organizations are urging public and philanthropic funders to use an expanded learning framework for evaluation and other methods of gaining knowledge from complex interventions, strategies, and initiatives.

Grantmakers for Effective Organizations (GEO) and the Council on Foundations are working together to nudge foundations to view organizational learning as a foundation function that is at least as important as assessing what they have accomplished, holding their grantees and themselves accountable, and developing and disseminating knowledge to their fields of interest.⁷³ GEO urges its members to use evaluation as a “powerful tool for improvement” rather than merely an accountability exercise. But its surveys indicate that grantmakers still see accountability as the primary purpose of evaluation.⁷⁴

According to studies from GEO, the foundations that do redefine their role to emphasize organizational learning for improvement are struggling to find new ways of collecting new kinds of evidence. Because they contend that “it’s about improvement, not just proof” and “it’s about contribution, not attribution,” they are not hung up on the difficulties of enrolling participants through random assignment, nor do they have to overcome all the barriers to finding appropriate “counterfactuals.” Because they have concluded that “it’s about going beyond the individual grant,” they are able to gain insights into the commonalities among different approaches to solving common problems—among their own grantees and, hopefully, among grantees of other funders addressing similar issues.⁷⁵

The Aspen Institute’s Roundtable on Community Change, meanwhile, reviewed 48 major community change efforts of the past two decades and analyzed changes in their approach to evaluation and learning. In its landmark 2010 report, *Voices from the Field Vol. III*,⁷⁶ the Roundtable pointed out that 15 years ago evaluators of community initiatives focused primarily on the challenge of causal attribution and debated whether experimental and quasi-experimental approaches could be used. Today, while major stakeholders continue to call for evaluations that attribute causality, many realize that experimental designs are unrealistic for community change enterprises. Many cutting-edge evaluators are likely to define their work as “contribution analysis” rather than “attribution analysis.” Evaluation is increasingly serving many purposes beyond summative assessment.

In her chapter in *Voices III*, “Evaluating and Learning from Community Change Efforts,” Prudence Brown observes that foundations that support community change (and help to design evaluations as well as invest in them) now give learning a more central place in their mission, goals, strategies, internal structures, and external partnerships.⁷⁷ Brown describes a movement toward shared evaluation frameworks, more realistic expectations for measuring impact, more attention to real-time learning, and new approaches to evaluating policy and systems change:

Funders and their partners articulate goals and strategies and specify measurable interim and long-term outcomes....They have developed a better understanding of the “attribution problem” and the difficulty of drawing a straight causal line between investments in community change and

^P This is the idea that the late MIT organizational theorist Don Schön described as “epistemological nihilism in public affairs,” the view that nothing can be known because the certainty we demand is unattainable.

specific outcomes. Evaluators have grown more aware of the multiple causal factors at play within the complex ecology of community change—and their clients have grown more interested in learning how to create change, not just proving that it has occurred.⁷⁸ As a consequence, they are using multiple methods and sources of data to “make a compelling case” that links the change effort with intended outcomes or lack thereof.

Evaluation in community change work has been increasingly viewed as a means to enhance real-time learning and decision-making, refine strategy, and institute midcourse corrections.⁷⁹

Two examples cited by Brown illustrate the kind of comparisons that can be made that shed a bright light on the workings of community initiatives without thrusting them into the Procrustean bed of experimental validation. *Vibrant Communities*, an initiative that operates in 12 Canadian communities to reduce poverty, developed an evaluation that, among other things, organizes results according to three levels of action (community capacity for poverty reduction, individual and household assets, and policy and systems change), each with four or five key indicators.⁸⁰ And in the *Chicago New Communities Program* (NCP), evaluators examine demographic changes as well as the nature, extent, and pace of change in such neighborhood indicators as crime rates, housing market activity, and commercial vitality. Their analysis can show how trajectories vary across NCP communities and how they compare to changes in selected non-NCP neighborhoods and for Chicago overall.⁸¹

A considerable amount of useful learning can be generated from complex interventions as part of the day-to-day management of work. One promising approach is the practice of developing a results framework, tracking progress toward those results, and using the data for “real-time” learning that continuously shapes and drives efforts. In this approach, leaders first specify target results in terms of the conditions of well-being—for children, families, or communities—that they plan to achieve. They then agree on both indicators of progress and the interim measures that can serve as reliable benchmarks because they are empirically linked to these long-term results (e.g., regular school attendance predicts improved academic achievement). By carefully and continuously monitoring progress toward the results, program leaders can infer the likely direction and size of their program’s effects and make mid-course corrections and move the work forward without having to wait for evaluation findings, which take time to produce.

One promising approach is the practice of developing a results framework, tracking progress toward those results, and using the data for real-time learning.

The process of managing to results can be particularly important when it is adopted by many partners and adhered to across multiple service systems. It then becomes a method of developing and reinforcing what Kania and Kramer refer to as “broad cross-sector coordination,” which is necessary to achieve any truly ambitious improvements in child or family well-being.

The growing thrust toward managing to results received a giant boost with OMB’s April 29, 2011 directive on implementing the Presidential Memorandum “Administrative Flexibility, Lower Costs, and Better Results.”⁸² The memo urges the heads of federal executive departments and agencies to find areas of operation where increased flexibility and greater collaboration could drive significant improvements in outcomes. They were asked to focus accountability on outcomes instead of processes, to define outcomes clearly and across jurisdictions, to share data and eliminate duplicative and unnecessary reporting, and to identify barriers that impede blending of funding streams.

OMB’s memo reflected careful listening to local leaders who have been implementing complex interventions that cut across systems and agency boundaries. For example, cities participating in the Annie E. Casey Foundation’s *Making Connections* initiative implemented over the past decade a version of management by results that is consistent with the intent of the new federal directive, although in the

case of *Making Connections* it applied to foundation as well as public funding. As part of efforts to simultaneously improve family earnings and income and ensure that young children grow up healthy, prepared to succeed in school, and successful in the early grades (known as a two-generation approach to achieving results), *Making Connections* established “results tables”—groups with representatives of schools, health and human service agencies, workforce and employment programs, city and county governments, neighborhood residents, and United Ways and local philanthropies. The results tables tracked the progress of their joint strategies and, using agreed-upon metrics, learned whether and how their new policies, programs, and practices were having the desired effects. Based on that knowledge, they expanded their strategies or changed course as indicated.

While each city’s effort was shaped by local leadership and priorities, the infrastructure that made managing to results possible had common characteristics, including:

- **A process for agreeing on a common set of results** across multiple sectors and partners. These shared results, along with a carefully developed set of indicators and interim benchmarks, remained the goals to which all partners devoted their efforts over many years.
- **Collaborative forums where partners came together regularly** to review data about progress toward the targets they had set, take stock of whether strategies were effective, and revise their plans as necessary.
- **Performance assessment tools and processes that simplified keeping score and heightened accountability.** Tools that proved effective included: (1) performance contracting, which set expectations for each partner’s contribution to achieving shared results; (2) an annual process of recalibrating year-to-year targets in light of prior progress and new circumstances, while holding fast to longer-term goals; and (3) “dashboard” reports that summarized metrics concisely and with a clarity that allowed them to be shared with community residents, agency executives, and funders.

Leaders in these cities, with support and technical assistance from the Annie E. Casey Foundation, developed processes that worked in local contexts. Other initiatives are developing the “next generation” of processes by which multiple stakeholders use an unwavering attention to results to manage their collective efforts over an extended period of time. A good example comes from the *Strive Partnership* described on page 18. The indicators *Strive* uses to monitor progress toward its seven target outcomes are easily understandable, reasonably similar (and therefore comparable) across states and school districts, produced by trusted sources, affordable to gather and report, consistently available over time, useful in the day-to-day work of *Strive* and its partners, and changeable by local actions.⁸³ *Strive* staff have found that by reviewing trends on these outcomes over time, they can highlight where they are having the greatest impact and where they need to focus more energy along the cradle-to-career journey.

The elements of the process of managing to results on a community-wide basis, as illustrated above, are still being identified and tested. Two things seem clear, however.

First, implementing this approach requires investment in several new community capacities. Current data and accountability systems are confined by the mandates and organizational structures of individual agencies and systems; they are not designed to be used across systems and sectors. Shifting to a more results-oriented, comprehensive, and integrative accountability system will require investments in: (1) ongoing capacity to gather, analyze, and process data needed for decisions across systems and sectors; (2) people with the skills and expertise to staff processes by which multiple partners review data and experience, learn from them, and chart their future course; (3) the capacity of

neighborhood residents to be influential leaders in this process; and (4) links to citywide, regional, and state-level policy and budget decision makers who control many of the resources needed to achieve results.

Second, developing this type of results process seems worth the investment. Done well, it has potential to:

- Focus the attention and commitments of multiple partners around a common set of results;
- Allow progress toward multi-year goals to be tracked in useful, meaningful increments;
- Help managers keep highly complex and multi-pronged efforts on track;
- Generate the “news” about achievement that can continue to inspire and motivate the many partners whose efforts must remain of high quality; and
- Provide the learning about “what works, today, in this community to change results” that is both the incentive and reward for continuing complex social interventions.

Collectively, a framework centered on learning, a commitment to managing by results and a pragmatic approach to evaluation, can alter the way we gather evidence from complicated initiatives that seek to achieve ambitious results for children, families and communities. Fig. 5 offers an example of how new and better ways of locating, assembling, and analyzing evidence could be applied to a federal initiative—in this case, *Promise Neighborhoods*. In the section that follows, we explore what it will take to strengthen the field’s ability to measure complex change in order to better support learning and accountability.

Fig. 5: A LEARNING FRAME IN *PROMISE NEIGHBORHOODS*

Promise Neighborhoods, a signature initiative of the Obama Administration and its Department of Education, is inspired by the approach and success of the Harlem Children's Zone. *Promise Neighborhoods* seeks to help other communities put in place "cradle-to-career" supports that help all children in a designated community grow up healthy, succeed in school, and complete a college education or secure another post-secondary credential.

Promise Neighborhoods has importance even beyond its status as a federal initiative. Many communities are planning and mobilizing to implement this approach, knowing that federal budget constraints and the tough competition for limited dollars mean they will have to implement it with funds from states, localities, and private sources. Viewed from this perspective, the *Promise Neighborhoods* pioneer sites' ability to generate learning and evidence about how all children can achieve its stipulated outcomes is a compelling, high-stakes venture. Their ability to do this (and the Department of Education's and philanthropic investment in helping them) will determine whether *Promise Neighborhoods* fulfills its potential as one of the most important vehicles for improving opportunities for children in low-income communities across the nation.

Fortunately, *Promise Neighborhoods*' design has set the groundwork for effective learning. The initiative incorporates a results frame: Guidance to prospective grantees requires that they pursue a core set of results for all children served, focusing on academic achievement and child health. The guidance also encourages grantees to develop community capacity to administer this type of complex intervention, and it is especially clear (and demanding) in requiring longitudinal data systems so that local initiatives will track individual children's progress and the initiative's influence on population-level change for all children in the target neighborhood, community, or tribal area.

With the stage set for learning in this way, *Promise Neighborhoods* is positioned to incorporate some of the approaches recommended in this paper. Specifically, learning will be enhanced if:

Promise Neighborhoods sites are encouraged and supported to put in place a process of results management as part of initiative implementation. Implementing a longitudinal data system, while challenging in itself, is just a first step. Additional steps are needed to develop the community processes and habits to harvest the data routinely, use data to inform and prompt decisions from a collaborative (or governing) body of stakeholders, and install a results process as the fulcrum for continuous learning that leads to course corrections. The Department can encourage investment in staffing and infrastructure for a community results process; allow time for this process to take hold; incorporate performance expectations for this process during the early years of the grants; and share knowledge and gather evidence about "what works" in terms of a results process, so that lessons learned by one community are shared by all.

The plans for evaluation are sequenced into stages that are appropriate for each phase of the initiative. The framework and methodologies that have emerged around "developmental evaluation" seem useful here. For instance, during the initial years the focus should be on *formative evaluation*, helping to put the eventual ingredients of an evaluation into place (e.g., an articulated theory of change) and on assessing whether the capacities for high-quality implementation (e.g., the pipeline of high-quality services, a result-based management infrastructure) are in place. While indicators of impact on children and families can and should be developed during these early years, expectations for change should be modest, reflecting the fact that new supports for children's success are still being assembled. Only after three or four years of implementation should the stakeholders expect to begin seeing changes in the major indicators of results.

Tools that advance site learning are provided to sites by the Department of Education, to be adapted locally. Local site leaders are in a good position to suggest what they need (and thus should be part of a co-design process of tools with the Department or providers of technical assistance). Possible tools include: (1) formats for data aggregation; (2) “dashboards” that promote local analysis and communication of key indicators and interim benchmarks; (3) guidance materials for staffing and running community results processes; (4) metrics to help communities self-assess their own growing capacities to implement *Promise Neighborhoods* pipelines; and (5) “learning questions” to help sites examine and benefit from their experiences. These learning tools would be as useful as recommendations for programmatic interventions—and, during the early years, probably more so (and more difficult for sites to find on their own).

The Department establishes and maintains a learning community among sites. The Department has already signaled its intent to do this, and private funding could be used to augment the public-sector investment. The people who implement complex interventions usually consider learning communities, or “communities of practice,” the most helpful form of learning and support. These networks require investment, however, to ensure that knowledge is captured and synthesized; otherwise they merely act as a source of mutual support, which is a useful but limited role.

The cross-site evaluation incorporates multiple methods. It will be especially important to: (1) document the effectiveness of implementation during the early years; (2) set benchmarks for the development of local capacity, so that sites’ readiness to achieve impact can be assessed regularly; and (3) incorporate into the eventual impact evaluation techniques that allow stakeholders to compare—persuasively—the interventions’ results with results achieved in other circumstances (even though experimental methodologies are unlikely to apply, except perhaps for limited program components).

SECTION 5: STRENGTHEN MEASUREMENT FOR ACCOUNTABILITY AND LEARNING

The paucity of good measurement tools is a formidable barrier to learning in real time, maintaining accountability, managing by results, continuously improving quality, and assessing impact. As agreement has grown among funders, policymakers, advocates, community residents, and practitioners that these functions are important, it has become increasingly apparent that high-quality, widely accepted, readily understood, user-friendly measures and indicators are not available where they are most needed. Here we discuss the reasons such tools are missing and suggest ways to support their development, availability, and common use.

ISSUE: *Few indicators are uniformly available to cities, neighborhoods, and other small geographic areas where place-based reforms operate*

SOLUTION: *Develop appropriate measures for smaller geographic units*

Child Trends, a Washington-based organization with arguably the most expertise in the country on measures and indicators of child, youth, and family well-being, warns about the scarcity of neighborhood- and community-level indicators in its introduction to “*Results and Indicators for Children: An Analysis to Inform Discussions about Promise Neighborhoods*.”⁸⁴ The authors note that local sites can develop local indicators to supplement federal and state data by examining administrative data from schools, municipal services, and health and social service agencies. Indeed, many communities have undertaken special-purpose surveys of residents or service providers to generate such data. However, the tasks of securing data-sharing agreements, organizing and managing the data, and collecting new data all require significant resources. And, because measures and definitions usually are not standardized—what constitutes child abuse, neglect, or school readiness varies widely from place to place, for instance, as does the type of crimes that get reported—these locally collected indicators are not easy to compare across sites or against national statistics.

We recommend several steps to address this issue:

- **Support and expand efforts such as the Urban Institute’s National Neighborhoods Indicators Partnership (NNIP).** NNIP recognizes the need for more systematic and comparable data in small geographic units and provides guidance for communities on how to maximize and blend the available administrative, survey and other data. With investment to continue NNIP’s influence, a much wider range of comparable indicators could become available.
- **Increase the array of CitiStat projects that mayors and other city officials have implemented to create reliable, on-going databases that measure local progress.** These systems represent the standing capacity to produce and use data needed if city-wide and neighborhood-based efforts are to thrive and endure.
- **Provide high-quality survey tools for place-based neighborhood initiatives.** Often, the cost of developing surveys from scratch is prohibitive for local communities, especially when the survey results cannot be reliably compared with those of other jurisdictions. An effort to provide communities with reliable survey instruments that they could use with confidence—sponsored by foundations or perhaps the U.S. departments of housing or education—could enable local leaders to obtain some types of neighborhood-specific information reliably and affordably.

- **Over time, work toward more intensive sampling of poor, minority children and families in national surveys, thereby generating locally reliable data.** This has already been done in a few cases, such as the Current Population Survey, a monthly survey conducted for the U.S. Bureau of Labor Statistics to track labor force characteristics, and the results have proved useful for local jurisdictions. Greater investment in sampling for these surveys, at a comparatively modest cost, would pay off by helping federal, state, and local initiatives better gauge their impact.

ISSUE: *What gets measured gets done*

SOLUTION: *Develop metrics to capture all critical areas of work*

Initiative leaders and front-line staff worry that easy-to-measure outcomes will take precedence over others that are equally or more compelling but harder to measure. Two issues are particularly important here. The first concerns which measures of progress and impact we use. For instance, will we judge a student’s reading progress solely by the ability to decode (something relatively easy to assess), without measuring reading comprehension? The second issue has to do with how broadly we think about what it takes to achieve progress and results. Measurement of community change typically focuses on the effects of interventions (for example, their influence on programs, systems or policies, and hopefully on their impact on children, families and neighborhoods) but pays less attention to documenting and measuring whether people and organizations also have developed the *capacities* needed to achieve and sustain those results—the “work behind the work” that is equally essential.

We recommend the following actions to keep this measurement challenge from undermining the focus on results:

- **In complex initiatives that measure child development and academic success over time, resist measuring only the most easily benchmarked characteristics.** Academic progress and, to a lesser extent, health conditions are essential measures of children’s development and more easily measured than social and emotional well-being, but it is important to measure the second set of changes too, even if the available indicators are imperfect.
- **Over the next five years, develop more uniform, field-tested indicators and measures of critical stages of healthy child development.** Ultimately, progress will not be possible unless new types of measurement are developed, tested, made reliable, and then widely used. Based on needs expressed by local leaders who have tried to implement pathways to developmental success for children, priorities for improving indicators could include creating: (1) measures for healthy child development in the first three years of life and then from that point to school entry, (2) measures for children’s readiness for school entry, (3) generally accepted measures for positive youth development, and (4) measures for young people’s readiness to enter and complete post-secondary education.
- **Develop measures of community capacity** that allow progress to be assessed as neighborhoods and local communities develop the strengths and functioning capacities needed to implement collective efforts to achieve results rather than more narrowly defined programmatic efforts.

ISSUE: *It takes time to achieve and document highly valued, significant results, but funders and policymakers often are impatient for just these results*

SOLUTION: *Create a range of interim measures and help funders and political leaders understand why they should expect and attend to these incremental signs of progress that predict long-term results*

The inherent conflict between the long time it takes to achieve documented results and the pressing needs of funders, political figures, and managers to demonstrate impact in the short term presents a major challenge to solving complex problems. Especially in this era of shrinking resources, people need evidence of progress, shared in a timely way, to judge whether an initiative is worth investing in. There is probably no greater challenge on the measurement horizon than finding the short-term indicators, benchmarks, milestones, or “leading indicators” that predict authentic long-term accomplishment.

There is no shortcut to resolving this problem, but we recommend the following actions to move toward a more appropriate approach:

- **Help funders understand the time frame required to achieve and sustain solutions to complex problems.** Almost inevitably, such efforts involve phases that begin with an absorption of new ideas and approaches; the development of coalitions to lead and support the community’s intensified progress toward specific results; small-scale testing and implementation of strategies; course corrections; and then the gradual implementation of strategies at fuller strength and greater scale. It would be useful to make this developmental arc more widely known to public and private funders so they can stage their expectations accordingly.
- **Improve communication from the evaluation community about the importance of, and approaches to, interim measures.** In addition to adjusting expectations of funders, it would be helpful for evaluators to provide guidance on appropriate interim measures to leaders of complex social interventions. These could be measures of incipient capacity or some other way of documenting that the components of a theory of change are being put in place. They could be measures of interim outcomes that predict more significant later outcomes, as steady attendance can predict school achievement, and parents routinely reading to young children predicts school readiness. Consistent messaging that interim measures are, in fact, the stepping stones to complex but powerful accomplishments would help ensure that community strategies have sufficient time to incubate, to be thoughtfully implemented—and thus have the greatest chance to succeed.

ISSUE: Programs are pushed to show they each have a direct impact, but the results that matter most require coherent, coordinated efforts by multiple agencies, stakeholders, and interventions

SOLUTION: Emphasize collective impact in results, accountability, and measurement frameworks by establishing shared results

There are few examples of success in resolving this tension between individual and collective impact. The push for greater accountability gives funders incentives to look for the “best” individual grantee rather than for combined strategies. As this happens less attention and investment go into helping initiatives that span multiple sectors and disciplines to build measurement systems and structures that support the achievement of organizational mission, show results and document accomplishments.⁸⁵ And yet it is essential to find common measures for multifaceted approaches if we hope to advance a common agenda for change.

We recommend the following actions to address this issue:

- **Keep the focus on results, especially those that require the collective action of multiple partners to produce better outcomes for children, families, or communities.** From this perspective, individual programs contribute to the conditions that produce results as *elements* of change but not as the total engine of change.

- **Develop strong results management frameworks and accountability systems, with regular measurement of progress and of each partner’s contribution to the overall result.** Shifting from program accountability to collective accountability (i.e., accountability shared by multiple people and organizations) need not diffuse responsibility, it just broadens it. However, this can only occur if the collective accountability framework is as clear and specific as an individual program assessment model would be.
- **Develop new measures associated with the ingredients of collective impact.** These are likely to focus on capacity measures, such as the joint commitments of agency partners, the co-investment of resources from many sources, and the common policy directions articulated and implemented by the stakeholders who are holding themselves accountability for outcomes.

To the mantra that fundamental change is about “relationships, relationships, relationships,” wise managers add, “but it doesn’t count unless it’s also about measurement and results.”⁸⁶ We believe that when community leaders have the tools and supports they need to jointly keep score, pursue agreed-upon results, and measure their progress in ways that make sense to them, their funders, and researchers, they will have reason to balance the focus on relationships with a strong orientation toward data, accountability, and performance. In fact, shared measures can enrich relationships, because they heighten accountability and make it easier to agree on specific methods and approaches with multiple partners.

Shared measures, together with results-based management, keep the focus on the outcomes sought for families and children when other factors threaten to pull attention away. Agreed-upon measures make it possible to examine the progress achieved for different populations, often raising equity issues in a way that forces people to deal with them. And shared measures also make visible, often for the first time, whether the theories of change behind human service and community development strategies actually produce results. This may not be a comfortable contribution, but it is always an important one.

None of the measurement challenges raised in this section can be resolved by individual local initiatives. And few will be overcome in a cost-effective way without considerable investment by the federal government and philanthropy in organizations that can take this work to a new stage. Vehicles do exist that have enough capacity to develop, disseminate, and facilitate the common use of appropriate measurement tools (as can be seen in the work of Child Trends and the Urban Institute’s National Neighborhood Indicators Partnership), but they have been given neither the charge nor the resources to do this task.

Philanthropy (initially) and the public sector (eventually) need to become more organized and intentional about developing appropriate data sources, indicators, and measures for complex interventions. This work cannot be done community by community, but it could be done by sector; in relation to a broad framework of target populations (e.g., children, youth, families); or, to start, in relation to specific measurement challenges.

An imaginative example of the latter can be seen in the Benchmarking Project, a collaboration funded by the Annie E. Casey Foundation through Public/Private Ventures (P/PV). Casey and PPV identified the most troubling challenges for the employment sector that arose from the absence of consistent workforce measures, compounded by the fragmentation of the workforce development system:

- Inconsistent definitions of outcomes across funding streams make it difficult for decision makers and practitioners to get a clear picture of overall program results.

- Diverse reporting requirements and the customized technology systems of different funders require duplicate and time-consuming data entry for practitioners, often without generating useful reports in return, and take significant time and energy to navigate, thus sapping frontline providers' capacity to use data for program improvement.
- Performance standards and comparisons that do not take program differences into account make it hard to know what constitutes "good" performance.

The Benchmarking Project is helping 159 organizations in three cities deal with these issues by moving toward more consistent definitions of performance measures, implementing new technology or adapting existing systems to allow funder and program databases to exchange information more easily, providing more useful reports for practitioners about local and state data trends, and offering more opportunities for program providers to learn from research and from their peers.⁸⁷

As the social reform sector gears up to strengthen its measurement and accountability capacity, concerns about *how* to measure must be joined with clarity about *what* to measure.

In a year-long review of foundation approaches to measurement and evaluation, the Aspen Institute's Program on Philanthropy and Social Innovation found that "grant-makers and their grantees tend to gather as much data as possible in search of evidence of impact." This frequently results in "expensive evaluations that are onerous for grantees and have often failed to yield the data needed in the time required for informed decision-making. As a result, the data have gone unused."⁸⁸ Instead, the Aspen Philanthropy Program recommends a "decision-based" approach to measurement, characterized by:

- A shared purpose of informing decision making and enabling continuous learning;
- A shared expectation that data will be gathered in a timely fashion and in a manner that does not place an undue burden on grantees; and
- A shared commitment to placing data gathered in the public domain so as to advance field-wide learning.

Moreover, as the Benchmarking Project and many other efforts to improve measures recognize, the technical task of developing the right indicators is not the only challenge. An equally necessary and even more durable change is to develop and maintain an internal organizational culture of continuous improvement. This point is echoed from many quarters. Child Trends, for instance, observes that:

[F]or some managers, the usefulness of information ends once they have met their funder's requirements. Such managers may be collecting performance *measures*, but they are not engaging in performance *management*.... Other managers regard data collection and analysis as ongoing parts of their jobs that enable them to develop a better understanding of their programs' strengths and challenges, which helps in identifying solutions to improve program operations.⁸⁹

Venture philanthropist Mario Morino, meanwhile, recommends "peer 'professional learning communities' for nonprofit leaders who are trying to influence their organizational cultures and establish systems for managing to outcomes." Resonating with our own recommendation, he also advocates "creating outcomes assistance centers, such as centers built around the remarkable Child Trends research on the outcomes that matter for children and youth."⁹⁰

CONCLUSION

The pragmatic, inclusive approach to evidence advocated in this paper can help all of us do a better job of understanding change and achieving positive outcomes for children, families, and communities. It can only take root and flourish, however, if several changes occur. Among the most important are:

1. Researchers and practitioners begin systematically to distill evidence and new knowledge from a wider and richer range of sources;
2. Federal agencies and private foundations allocate funds in ways that recognize, incentivize, and support development of a broader body of evidence through rigorous capture of the learning generated in real time by innovative initiatives;
3. Communities develop or acquire the tools necessary to capture and learn from their experience; and
4. Evaluation techniques that respond to the multi-layered strategies that make up systems change efforts and complex community initiatives become widely used.

We conclude by describing these changes and identifying how foundations, federal officials, and local leaders can begin to put them into practice.

Step 1: Support knowledge collection, analyses, and evidence syntheses that yield a more complete body of evidence, leading to improved outcomes for children, families, and communities.

Early steps toward putting together the evidence now available are already under way. For example, several evidence-based practice registries have been created with funding from a combination of federal agencies and foundations. Such lists are increasingly visible in the fields of education, health and mental health, and other human services. However, lists of individual programs that have been assessed by experimental methods represent only a small fraction of the rich evidence base required to support the kinds of reforms that will significantly improve outcomes for children and families.

More inclusive syntheses of knowledge and evidence that combine findings of programmatic evaluations with findings from other kinds of research, theory, and practice will provide the kind of information that enriches the policy and practice decisions of tomorrow. One example, created by groups operating under the auspices of Harvard University's Center on the Developing Child, is described on page 7 of this paper. Other examples come from efforts to extract the commonalities among programs aimed at similar outcomes; perhaps the most ambitious are the ones we described on pages 14-15, conducted by Child Trends and by Georgetown University for the Department of Justice. These "evidence syntheses" gather, distill, communicate, and disseminate a full range of useful knowledge about the components of effective interventions that can be extracted at a level deeper than the identification of model programs, and therefore provides more "actionable" guidance to policymakers and practitioners.

To move in this direction, public and private funders should:

- **Support this type of synthesis and knowledge development in all the outcomes areas** related to healthy child development, educational achievement, family functioning, and community development; and

- **Over time, consider institutionalizing this process by identifying and supporting “knowledge centers”**—based in universities, public policy centers, or consulting groups—where developing, communicating, and updating evidence syntheses would be an on-going process.

The product of this change would be an enriched, ever-current knowledge base available to the policymakers and others in the field who must make daily choices about how best to achieve outcomes for children, families and communities.

Foundation funding could be especially effective in the near term by financing prototypes of this type of synthesis. In doing so, foundations would continue their leading role in arguing for high standards of evidence while simultaneously seeking the benefits of a richer definition of evidence. Eventually, however, federal funders should also invest in this area of knowledge development. The policy/research arms of the federal Departments of Housing and Urban Development, Education, Health and Human Services, and Justice, for example, are ideally positioned to commission rigorous academic analyses and evidence syntheses that would yield the expanded evidence base necessary for the next wave of change. And federal funders may be in a better position than foundations to create this process as a permanent resource for the field.

Step 2: Ensure that state- and community-level initiatives—whether or not they include innovations—can generate new evidence by tracking progress against desired outcomes, documenting progress, adjusting strategies in response to feedback, and capturing what they have learned in a rigorous way.

This recommendation reflects an assumption explained in detail earlier in this paper: The next wave of evidence needed to improve outcomes for children, families, and communities will come from learning (a) how complex intervention strategies (not just programs) interact with community conditions, opportunities, policies, and practices and (b) what impact these interactions have on the well-being of children and families. Gathering evidence about this more intricate array of interventions means requiring that they have a structured process to define desired outcomes, articulate the pathway to reach those outcomes, track progress, learn in real time why progress does or doesn’t occur, and document learning in a way that can be shared and applied.

By highlighting these components of innovative and complex interventions with the same intensity that they now highlight evidence-based practices and programs, federal and/or foundation investors would cause every effort they fund to generate new evidence.

Step 3: Accelerate the development of the tools and capacities that will help local communities generate new knowledge at greater scale.

We direct this recommendation to foundation funders in particular and, to a lesser extent, federal agencies. As funders establish higher expectations for developing real-time evidence, states and communities will need help in developing that capacity. This requires more than the ability to “evaluate” in the traditional sense of the word. What is needed is the capacity to manage complex interventions and innovations so that evidence development is “hard-wired” into, and inseparable from, day-to-day operations. At a minimum, greater investment will be necessary in the tools, infrastructure, and processes that allow local stakeholders to:

- **Launch and sustain a collaborative process among multiple partners** (e.g., schools, social service providers, health agencies, businesses, housing authorities, community development organizations, and local, state, and federal agencies) to identify, pursue, and achieve broadly accepted results;
- **Gather and use data for effective decision making and to track progress** against results and identify the reasons progress was or was not made; and
- **Assess and evaluate the factors that contribute to success** when it occurs or can help explain the failure to achieve desired outcomes.

A limited number of communities have successfully implemented the type of learning processes envisioned here. We now need a sustained effort to develop the tools, training, and technical assistance that could allow many more communities to put these capacities reliably in place.

Foundations are ideally positioned to help scale up these community capacities. In fact, the concept of “community capacities”—that is, the wherewithal to accomplish broad-scale change at the community level—has largely emerged from foundation-funded initiatives. To enable these capacities to be understood more completely and developed at a larger scale, national foundations should invest over multiple years to:

- **Refine what is currently known** about capacity-building strategies;
- **Develop tools** that can help others implement them;
- **Test metrics** that allow implementation of capacities to be reliably tracked by federal funders and others; and
- **Support technical assistance** to help communities with this task and to continue building the state of the art.

Step 4: Continue to expand the menu of available evaluative techniques and do a better job of matching them to different types of interventions.

This recommendation is for funders and the evaluation community, working together. On the funders’ side, the federal role is particularly important here. The large-scale initiatives of the Obama Administration have capacity to set the tone for evaluative practice for years to come. As evaluation designs are developed for complex initiatives (such as *Promise Neighborhoods*, *Choice Neighborhoods*, and the *Investing In Innovation Fund*), funders should encourage their evaluators to identify how they will match their various evaluation techniques to the challenges they will face in accumulating evidence from the work on the ground. Evaluators should be required to use multiple techniques to generate the fullest possible amount of knowledge from these initiatives and to encourage the development of shared outcomes among initiatives aimed at solving similar problems. While experimental designs may be appropriate to gauge the effects of some parts of the interventions on some children, other kinds of evaluation techniques will be essential to understand how these large-scale changes are best implemented in complex communities and systems, in difficult economic times with fierce resource constraints, and in ways that respond to very different cultures and community systems.

In short, the answer to the core questions of “what worked, for whom, to what extent, how and why?” cannot possibly be answered through a single experimental design. A different “gold standard” will have to be recognized and valued.

For federal funders of evaluations for these initiatives, the challenge will be to:

- **Develop evaluation specifications that encourage innovative evaluation techniques;** if experimental designs are required for certain interventions or subsets of target populations, augment these designs with other techniques; discourage the use of randomized experiments except when the need for certainty outweighs their high costs in time, money, and narrowness of findings.
- **Select evaluators who have a track record of evaluating complex, multifaceted interventions** with a combination of formative and qualitative evaluation techniques.

The evaluation community can also help move toward a more inclusive approach to evidence.

No one step alone will pave the way for a more inclusive and effective approach to gathering, analyzing, and using evidence.

However, by opening the door to multiple knowledge sources, the combined influence of the four steps suggested above would go a long way toward helping us all learn more and, consequently, achieve more.

The debate about standards of evidence is neither academic nor trivial. It has a significant impact as positions on this issue shape policies and investment strategies related to the persistent challenges of impoverished neighborhoods, failing schools, joblessness, and other factors that leave so many children and families behind. How we as a nation deal with issues of evidence will shape the nature of social innovation—what is and what is not allowed, promoted, and incentivized—for years to come.

How we as a nation deal with issues of evidence will shape the nature of social innovation—what is and what is not allowed, promoted, and incentivized—for years to come. The costs of not using all available evidence to develop more nuanced and powerful strategies for change are too high to keep paying.

We make a grave mistake if we approach evidence too narrowly. Too much potential for innovation, and for improved outcomes will be lost if we continue to apply a too-limited definition of what counts as credible evidence and privilege only a select few research methods to understand the complex, multifaceted strategies required to address poverty, inadequate education, joblessness, and years of disinvestment in low-income communities. As these problems worsen, as disparities of income and wealth grow ever larger, and as large segments of the nation’s population—especially its young people—seem to slip further away from genuine opportunity, the costs of not gathering, analyzing, and using all available evidence to develop more nuanced and powerful strategies for change are too high.

It is for that reason that we urge a more inclusive approach to evidence. We hope that this paper contributes new thoughts to the discussion and moves us toward some common ground on these issues.

ENDNOTES

- ¹ Drucker, P.F. (1980). *Managing in Turbulent Times*. New York: Harper & Row.
- ² http://www.childtrends.org/Files/Child_Trends-Lifecourse_Interventions.pdf
- ³ Telephone Conversation with the Nurse Family Partnership Central Office, March 22, 2011
- ⁴ Pecora, P., Kessler, R., Downs, A.C., English, D., White, J., and Heeringa, S. (March 2007). "Why Should the Child Welfare Field Focus on Minimizing Placement Change as Part of Permanency Planning for Children?" Paper presented at the California Permanency Conference.
- ⁵ Center on the Developing Child at Harvard University. (April 2007). "A Science-Based Framework for Early Childhood Policy: Using Evidence to Improve Outcomes in Learning, Behavior, and Health for Vulnerable Children"; and Center on the Developing Child at Harvard University. (2010). "The Foundations of Lifelong Health are Built in Early Childhood." <http://www.developingchild.harvard.edu>.
- ⁶ Summary of the Patient Protection and Affordable Care Act (Health Care Reform Act H.R. 3590-216), <http://www.doh.wa.gov/cfh/micah/hvna/sumryneedassess/default.htm>
- ⁷ Nurse Family Partnership, <http://www.nursefamilypartnership.org/proven-results>
- ⁸ Daro, D., Dodge, K.A., Weiss, H., and Zigler, E. (April 21, 2009). Letter to the President, <http://mainecgc.org/Daro%20Dodge%20Weiss%20Zigler%20Comments%20on%20Home%20Visiting%20Proposal.pdf>
- ⁹ Shonkoff, J. (January/February 2010). "Building a New Biodevelopmental Framework to Guide the Future of Early Childhood Policy." *Child Development*, Vol. 1 No.1.
- ¹⁰ Hoffmann, E. (April 2010). "Extending Home Visitation Programs to Family, Friend, and Neighbor Caregivers and Family Child Care Providers," <http://www.CLASP.org/admin/site/publications>
- ¹¹ Halle, T., Forry, N., Hair, E., Perper, K., Wandner, L., Wessel, J., and Vick, J. (2009). "Disparities in Early Learning and Development: Lessons from the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B)." Washington, DC: Child Trends.
- ¹² Daro, D. and Dodge, K. (2010). "Strengthening Home-Visiting Intervention Policy." *Investing in Young Children*. Brookings Institution and NIEER.
- ¹³ *Ibid.*
- ¹⁴ U.S. Department of Education. (October 2009). *Investing in Innovation Fund*, <http://www2.ed.gov/programs/innovation/factsheet.html>.
- ¹⁵ Alyson Klein, "Event Aims to Give 'i3' Competition Momentum," *Education Week*, January 19, 2011
- ¹⁶ Kumanyika, S.K. and Institute of Medicine Committee on Evidence Framework for Obesity Prevention Decision Making. (April 2010). *Bridging the Evidence Gap in Obesity Prevention: A Framework to Inform Decision Makers*. Washington, DC: National Academies Press, www.nap.edu.
- ¹⁷ Kelly, T. (January 17, 2011). Personal communication.
- ¹⁸ Lipsey, M.W., Howell, J.C., Kelly, M.R., Chapman, G., and Carver, D. (December 2010). "A New Perspective on Evidence-Based Practice." *Washington, DC: Georgetown University Public Policy Institute's Center for Juvenile Justice Reform*.
- ¹⁹ For a full listing of topics see http://www.childtrends.org/_catdisp_page.cfm?LID=B1F74E5F-7B8D-466C-B7CF2E9CC378EBDA.
- ²⁰ Emig, C. (February 25, 2011). Personal communication.
- ²¹ Grantmakers for Effective Organizations and Council on Foundations. (2009). "Evaluation in Philanthropy: Perspectives from the Field." Authors.
- ²² Moore, K.A., *et al.* (2006). "Program Implementation: What Do We Know?" Washington, DC: Child Trends.
- ²³ Clinton, W.J. (May 14, 1993). "Remarks by the President in Blue Ribbon Ceremony."
- ²⁴ Clinton, W.J. (January 19, 1993). Remarks by the President-elect, meeting with governors.
- ²⁵ Obama, B. (July 18, 2007). "Remarks of Senator Barack Obama: Changing the Odds for Urban America." Washington, DC: <http://www.barackobama.com>.
- ²⁶ Shonkoff, J. (2010). "Building a New Biodevelopmental Framework to Guide the Future of Early Childhood Policy." *Child Development*, Vol. 1 No.1.
- ²⁷ Lipsey, M.W., Howell, J.C., Kelly, M.R., Chapman, G., and Carver, D. (2010). "A New Perspective on Evidence-Based Practice." Georgetown University: Public Policy Institute Center for Juvenile Justice Reform.
- ²⁸ Knitzer, J., and Cooper, J. (2006). "Beyond Integration: Challenges for Children's Mental Health."

-
- ²⁹ Bradach, J.L. (Spring 2003). "Going to Scale: The Challenge of Replicating Social Programs." *Stanford Social Innovation Review*.
- ³⁰ Miller, W.R., Sorensen, J.L., Selzer, J.A., and Brigham, J.S. (2006). "Disseminating Evidence-Based Practices in Substance Abuse Treatment: A Review with Suggestions." *Journal of Substance Abuse Treatment*, No. 35, 25–39.
- ³¹ Summerville, G. and Raley, B. (2009). "Laying a Solid Foundation: Strategies for Effective Program Replication." Philadelphia, PA: Public/Private Ventures.
- ³² Bradach, J.L. (April 2003). "Going to Scale: The Challenge of Replicating Social Programs." *Stanford Social Innovation Review*.
- ³³ Public/Private Ventures. (2011). "Priorities for a New Decade: Making (More) Social Programs Work (Better)." Authors.
- ³⁴ Kania, J. and Kramer, M. (Winter 2011). "Collective Impact." *Stanford Social Innovation Review*.
- ³⁵ Summerville, G. and Raley, B. *op. cit.*
- ³⁶ *Ibid.*
- ³⁷ Kubisch, A., Auspos, P., Brown, P. and Dewar, T. (2010). "Voices from the Field Vol. III: Lessons and Challenges from Two Decades of Community Change Efforts." Washington, DC: Aspen Institute Roundtable on Community Change.
- ³⁸ *Ibid.*
- ³⁹ Burns, T. "Response Essay." Kubisch, A., *et al.*, *op. cit.*
- ⁴⁰ Public/Private Ventures (2011), *op. cit.*
- ⁴¹ Rittel, H. and Webber, M. (n.d.) "Dilemmas in a General Theory of Planning." *Policy Sciences*, 4, 1973, 155-169.
- ⁴² Kania, J. and Kramer, M., *op. cit.*
- ⁴³ Kumanyika, S.K. and Institute of Medicine, *op. cit.*
- ⁴⁴ LaMarche, G. (April 19, 2008). "Philanthropy Can Be Made to Measure." *Financial Times*.
- ⁴⁵ Kania, J. and Kramer, M., *op. cit.*
- ⁴⁶ Flower, J. (July-August 1995). "A Conversation with Ronald Heifetz: Leadership Without Easy Answers." *The Healthcare Forum Journal*, Vol. 38, No. 4; and Heifetz, R.A., Kania, J., and Kramer, M. (December 2004). "Leading Boldly." *Stanford Social Innovation Review*.
- ⁴⁷ Kumanyika, S.K. and Institute of Medicine, *op. cit.*
- ⁴⁸ <http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy>
- ⁴⁹ It is said that his formulation was that "only randomized trials are bullshit resistant."
- ⁵⁰ Goldberg, S. (2009). *Billions of Drops in Millions of Buckets: Why Philanthropy Doesn't Advance Social Progress*. Hoboken, NJ: John Wiley & Sons, Inc.
- ⁵¹ *Ibid.*
- ⁵² Peterson, P. (December 8, 1999). "Can We Make Education Policy on the Basis of Evidence? What Constitutes High Quality Education Research and How Can It Be Incorporated Into Policymaking?" *Brookings Press Forum*.
- ⁵³ Hollister, R.G. and Hill, J. (1995). "Problems in the Evaluation of Community-Wide Initiatives." New York: Russell Sage Foundation, <http://epn.org/sage/rsholl.html>.
- ⁵⁴ Houstoun, F. O'C. (February 2, 2011). "Knowing What Works: Evaluating Evidence-Based Programs." *Governing*, <http://www.governing.com>.
- ⁵⁵ Berwick, D.M. (2005). "Broadening the View of Evidence-Based Medicine," *Qual Saf Health Care* 2005;14:315-316 doi:10.1136/qshc.2005.015669; and Berwick, D.M. (December 11, 2007). "Eating Soup with a Fork." Nineteenth Annual National Forum on Quality Improvement in Health Care, Orlando, FL.
- ⁵⁶ *Ibid.*
- ⁵⁷ U.S. Government Accountability Office. (November 2009). "Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions." GAO-10-30, <http://www.gao.gov>.
- ⁵⁸ Kramer, M., quoted in Goldberg, S., *op. cit.*
- ⁵⁹ Berwick, D.M. (2007), *op. cit.*
- ⁶⁰ Fiester, L., with Smith, R. (August 2007). "Learning While Doing the Three Rs: Roles, Results, and Relationships." A Making Connections Working Paper. Annie E. Casey Foundation: Working document, pp. 6-8.
- ⁶¹ *Ibid.*
- ⁶² Eckhart-Queenan, J. and Forti, M. (2011). "Measurement as Learning: What Nonprofit CEOs, Board Members, and Philanthropists Need to Know to Keep Improving." The Bridgespan Group, www.bridgespan.org.
- ⁶³ LaMarche, G., *op. cit.*
- ⁶⁴ Kumanyika and Institute of Medicine, *Bridging the Evidence Gap in Obesity Prevention*

-
- ⁶⁵ Kumanyika, S.K. and Institute of Medicine, *op. cit.*
- ⁶⁶ Fiester, L. (2011). "Plain Talk About Teen Sex: Evaluating a Community-Based Effort to Change Adolescent Behavior Through Adult Communication." Baltimore, MD: The Annie E. Casey Foundation.
- ⁶⁷ U.S. Government Accountability Office. (November 23, 2009). "Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions." Washington, DC: Author.
- ⁶⁸ Harlem Children's Zone. (n.d.). "From Cradle Through College: Using Evidence-Based Programs to Inform a Comprehensive Pipeline." http://www.emhd.us/documents/From_Cradle_through_College.pdf
- ⁶⁹ Kubisch, A., Auspos, P., Brown, P., and Dewar, T., *op. cit.*
- ⁷⁰ Kania, J. and Kramer, M. *op. cit.*
- ⁷¹ Kubisch, A. *et al.*, *op. cit.*
- ⁷² Kumanyika, S.K. and Institute of Medicine, *op. cit.*
- ⁷³ Hunter, D. (2006). "Daniel and the Rhinoceros." New York: Edna McConnell Clark Foundation.
- ⁷⁴ Grantmakers for Effective Organizations and Council on Foundation. (December 15, 2009). "Evaluation in Philanthropy: Perspectives from the Field." Authors.
- ⁷⁵ *Ibid.*
- ⁷⁶ Kubisch, A., *et al.*, *op. cit.*
- ⁷⁷ Brown, P. (2010). "Evaluating and Learning from Community Change Efforts." In Kubisch, A., *et al.*, *op. cit.*
- ⁷⁸ Behrens, T.R. and Kelly, T. (2008). "Paying the Piper: Foundation Evaluation Capacity Calls the Tune." *New Directions for Evaluation* 119 (Autumn): 37-50; and Westley, F., Zimmerman, B., and Patton, M.Q. (2006). *Getting to Maybe: How the World Is Changed*. Toronto: Random House Canada.
- ⁷⁹ Bailey, T., Jordan, A., and Fiester, L. (2006). "A Framework for Learning and Results in Community Change Initiatives: Imagine, Act, Believe." Baltimore, MD: Annie E. Casey Foundation, www.aecf.org; Walker, G. (2007). "Midcourse Corrections to a Major Initiative: A Report on the James Irvine Foundation's CORAL Experience." Los Angeles: James Irvine Foundation, <http://www.irvine.org>.
- ⁸⁰ Leviten-Reid, E. and Cabaj, M. (2008). "Learning and Evaluation for Vibrant Communities Trail Builders: The Pan-Canadian Process." <http://www.tamarackcommunity.ca>.
- ⁸¹ Chicago New Communities Program, <http://www.newcommunities.org/whoweare>.
- ⁸² Jacob J. Lew, Director, Office of Management and Budget, April 29, 2011, Memorandum for the heads of executive departments and agencies, "Administrative Flexibility, Lower Costs, and Better Results for State, Local, and Tribal Governments"
- ⁸³ <http://www.strivetoegether.org/wp-content/uploads/2010/11/2010StriveReportCard.pdf>.
- ⁸⁴ Moore, K.A., Murphey, D., Emig, C., Hamilton, K., Hudley, A., and Sidorowicz, K. (2009). "Results and Indicators for Children: An Analysis to Inform Discussions About Promise Neighborhoods." Washington, DC: Child Trends.
- ⁸⁵ Ebrahim, A. and Rangan, V.K. (July 9, 2010). "The Limits of Nonprofit Impact: A Contingency Framework for Measuring Social Performance." *Working Knowledge*, Harvard Business School.
- ⁸⁶ Farrow, F. "Response Essay." In Kubisch, A.C., Auspos, P., Brown, P., and Dewar, T. (2010). *Voices from the Field, Vol. III: Lessons and Challenges from Two Decades of Community Change Efforts*. Washington, DC: The Aspen Institute, pp. 68-72
- ⁸⁷ Miles, M., Maguire, S., Woodruff-Bolte, S., and Clymer, C. (November 2010). "Putting Data to Work: Interim Recommendations From The Benchmarking Project." Philadelphia: Public/Private Ventures.
- ⁸⁸ Wales, J. (September 13, 2010). "Metrics that Matter: Venture Philanthropy Pioneer and Aspen Philanthropy Group Draw Similar Conclusions." <http://www.HuffingtonPost.com>.
- ⁸⁹ Walker, K.E. and Moore, K.A. (January 2011). "Performance Management and Evaluation: What's the Difference?" Washington, DC: Child Trends.
- ⁹⁰ Morino, M. (July 8, 2010). "Social Outcomes: Lifting Sights, Changing Norms." Washington, DC: Venture Philanthropy Partners.