

**Broader Evidence for Bigger Impact**  
By Lisbeth B. Schorr

Stanford Social Innovation Review  
Fall 2012

Copyright © 2012 by Leland Stanford Jr. University  
All Rights Reserved

---

# Broader Evidence for Bigger Impact

DEFINING CREDIBLE EVIDENCE HAS POLARIZED INTO TWO CAMPS—THE  
EXPERIMENTALISTS AND THE INCLUSIONISTS—WHICH MUST BE BROUGHT  
TOGETHER TO TACKLE SOCIAL PROBLEMS EFFECTIVELY.

BY LISBETH B. SCHORR

ILLUSTRATION BY BEN WISEMAN

**N**o one questions President Obama's insistence that public funds should go to social programs that work and not to those that don't. The controversy is about how we know what works, and the types of evidence that prudent investors should consider credible. The answers to these questions, like so much else in today's discourse, have become polarized into two camps.

The Experimentalists assert that trustworthy evidence comes only out of experimental evaluations, where participants that get the pill and the control group that gets the placebo are randomly selected. When Peter Orszag directed the White House Office of Management and Budget (OMB), he explained that the one guarantee against funding what doesn't work is experimental proof.<sup>1</sup> The story goes that his formulation was that "only randomized trials are bullshit resistant."

The Inclusionists contend that although appropriately applied randomized trials are uniquely valuable because they provide *proof*, any knowledge base that relies only on experimental evaluations is too narrow to be useful. The Inclusionists agree with the Experimentalists that evidence must be rigorously collected and analyzed, but insist on drawing on a richer array of evidence that comes out of practice and non-evaluation research. They contend that evidence-based does not have to mean experiment-based.

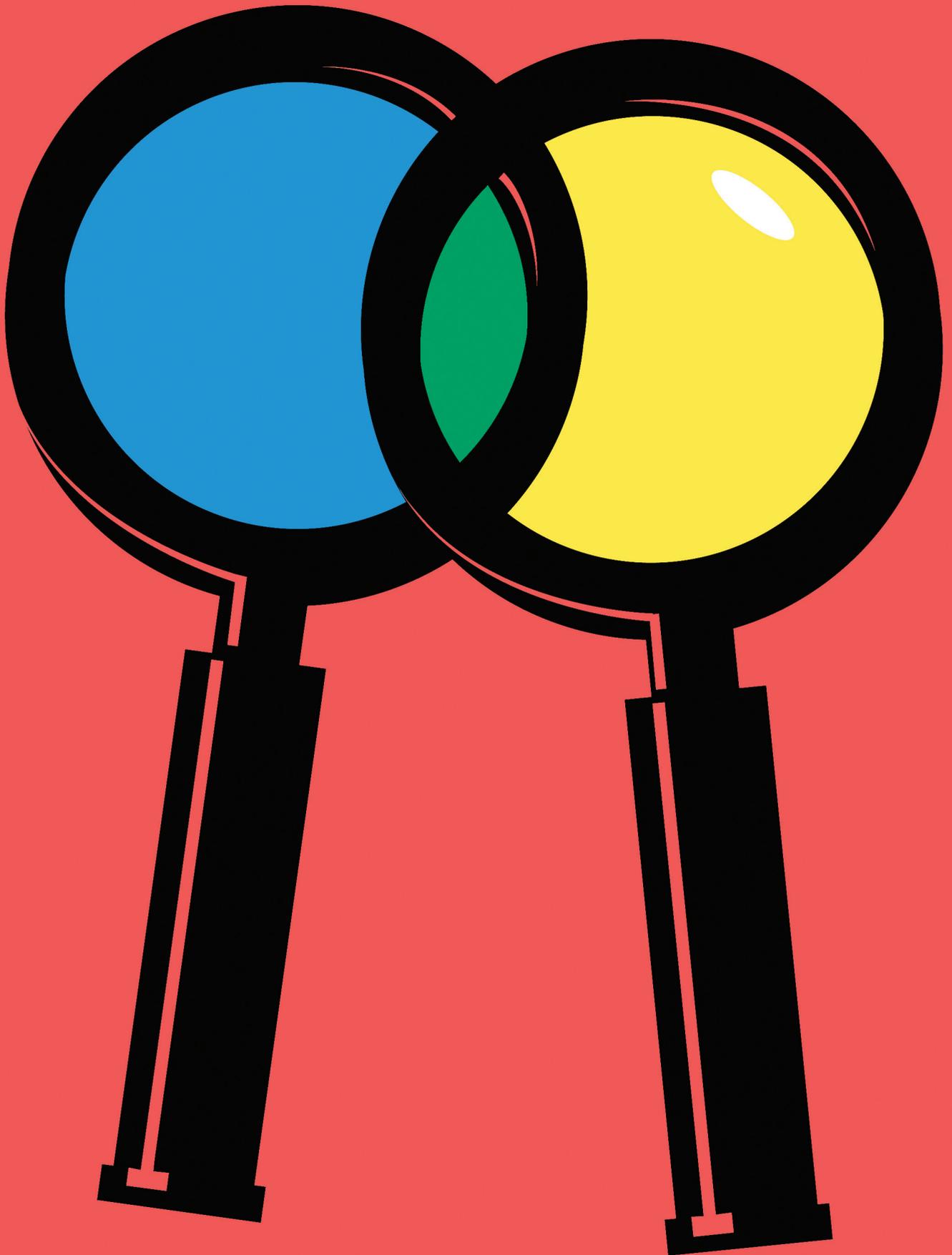
The stakes in this debate are high and getting higher, and the issue of determining what works is becoming ever more contentious

amid cutbacks in government spending, more constrained philanthropic giving, and an atmosphere of wholesale diminution of trust in our institutions and leaders.

But there are signs of progress in the search for common ground. In May, for example, the OMB issued a new memo on evidence and evaluation.<sup>2</sup> Although the directive reiterated and expanded its commitment to basing budgeting decisions on evidence, with an emphasis on experimental designs, it opens the way to the kind of reconciliation envisioned in this article. It asks agencies to suggest innovative uses of evidence and evaluation, and to propose new evaluations.

This is good news for those working to reduce widespread racial, income, and geographic disparities in health, education, child welfare, justice, and employment. To get better results we cannot rely solely on the interventions that were proven successful in the past, especially given the enormous advances of the last several decades in science, in our understanding of human behavior, and in our insights into the institutional obstacles that can prevent real change.

Rather, we need the transformative interventions that lead to what Dr. Jack Shonkoff, director of Harvard University's Center on the Developing Child, and his colleagues at the center call *breakthrough impacts*.<sup>3</sup> These types of interventions will become more likely when we find common ground between the Experimentalists and Inclusionists. Then past successes will become the starting point, not the final destination.



In this article I review the emergence of randomized trials as a way of understanding the impact of certain categories of interventions, and the reasons that experiments are not the only credible source of information about what works. I describe real-world efforts to generate and use different kinds of evidence, and I conclude with how it might be possible to attenuate the divisions between the two camps. If my argument occasionally tips the scale on behalf of the Inclusionists, it is because the prevailing wisdom has bolstered the Experimentalists for the past several decades.

## DETERMINING WHAT WORKS

**R**andomized trials to identify what works first emerged in psychology in 1885 and in agriculture in 1925. The first published randomized clinical trial (RCT) appeared in a 1948 paper on Streptomycin treatment of tuberculosis. The most dramatic impact made by early RCTs was on the treatment of breast cancer by radical mastectomy, developed by Dr. William Halsted in the early 1890s. Surgeons performed the Halsted procedure for more than 80 years, until new scholars subjected the procedure to randomized trials—an evaluation method unknown to Halsted. They found that radical mastectomy had no advantage over less intrusive treatments.<sup>4</sup> By the late 20th century, RCTs were recognized as the standard method for “rational therapeutics” in medicine.

In the mid-1960s, randomized trials began to move from medicine to the assessment of social programs, but the evidence from evaluation played a minor role alongside ideology and politics in determining the fate of social programs. Today the balance has shifted. Politics still frequently trumps rational analysis, but the era when a compelling vision, good intentions, and widespread local support could be counted on to be enough didn't last. The shine wore off antipoverty programs in the face of the Vietnam War, competing demands for funding, occasional failures and scandals, and ultimately President Reagan's campaign to shrink government. Calls for evidence of effectiveness grew, usually in the form of demands for evaluations using randomized experiments.

Social scientists developed experimental techniques to assess negative income tax experiments, employment and training programs, and welfare reform. Many cast envious eyes at the respect RCTs were accorded in medicine. As legislators began to mandate the use of experimental design as a condition of funding and foundation boards demanded objective evidence of results, a substantial research industry grew. Web-based clearinghouses appeared, featuring lists of interventions certified by experimental trials.

No one seems to know why experimental evidence is demanded (and provided) at some times and not others, or for some kinds of interventions and not others. Efforts to discern a pattern suggest that evaluation resources are most likely to go to new programs and policies being advanced when calls for proof of impact are loudest (such as when home visiting became part of the Affordable Care Act), to interventions considered circumscribed enough to be amenable

to experimental assessments (such as the well-defined Success for All in education), and to interventions targeting populations not universally considered “deserving” (such as poor people).

A review of collections of proven programs and the domains highlighted by major evaluation organizations suggest that current pressures for experimental evaluation focus on early childhood, K-12 education, teen pregnancy prevention, disposition of juvenile offenders, children's health and mental health, child welfare, youth development, family support, and training of marginal workers. In contrast, evaluation funds have not generally gone where public policies were firmly established (such as farm subsidies, military procurement, or the home mortgage interest deduction). A 2009 OMB initiative to strengthen impact evaluations focused especially on “social, educational, economic, and similar programs whose expenditures are aimed at improving life outcomes for individuals.”<sup>5</sup>

Contentions that evaluating complex social programs was not like testing new drugs were often seen as “unscientific,” as were warnings that most promising interventions to rescue inner-city schools, strengthen families, and rebuild neighborhoods could not be experimentally assessed.

Both producers and consumers of evaluation research seemed to have been intimidated into accepting narrow definitions of evidence as the only evidence worth having. And of course it was hard to resist the idea that you can actually rely on proof from incontrovertible numbers to identify the social interventions worth investing in, freeing funders from having to make fallible judgments. As Swarthmore College economist Robinson Hollister explained, 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.”<sup>6</sup>

And yet there has been a slowly growing recognition that no single approach to evaluation would allow us to learn enough from past interventions or enable us to predict the success of future efforts. The costs of RCTs seemed excessive and the results took too long to arrive. Interventions proven effective with RCTs didn't have the same successes when they were scaled up. Experimental evaluations didn't provide enough information about *how* the work affected outcomes, or about the political, economic, and regulatory contexts that could spell success or failure.

The interventions that seemed most likely to result in significant improvements in outcomes were hardest to assess in traditional ways. They were complex, had to be adapted to a variety of cultures and populations, required reforms of institutions, policies, and systems, and were continually evolving in response to changes in context, lessons learned, and advances in knowledge. Challenges to the conventional wisdom about what constituted credible evidence of effectiveness even began to come from medicine. The medical reformer Dr. Donald Berwick wrote that the RCT is a powerful, perhaps unequaled, research design, but only to explore the efficacy of those components of practice that are conceptually neat and have a linear, tightly coupled, causal relationship to the outcome of interest.<sup>7</sup>

Today, all those committed to achieving reforms that result in breakthrough impacts agree that whatever method they use to learn about their work, it must enable them to continuously make interventions more effective, guide the selection and design of interventions to be implemented or scaled up, and demonstrate that their work is

LISBETH B. SCHORR is a lecturer in social medicine at Harvard University and a senior fellow at the Center for the Study of Social Policy. She was the founding co-chair of the Aspen Institute's Roundtable on Community Change and has held leadership positions in many national efforts on behalf of children and youth, including the National Center for Children in Poverty, City Year, the National Academy of Science's Board on Children and Families, and the Foundation for Child Development.

improving lives and neighborhoods. But each camp proposes a very different basis for decision-making and action. The Experimentalists offer lists of programs that have been proven effective with experimental evaluations and contend that future action should consist of replicating these programs with fidelity to the original. Thus OMB declared in 2009 that the bulk of new federal funds would be made available for interventions that had been proven to work in the past.<sup>8</sup> The Inclusionists point out that the Experimentalists don't see the high price paid for certainty when funders and policymakers insist that funds should go only to interventions shown to be evidence-based, using a narrow definition of credible evidence. Reformers worry that, given the nature and magnitude of the problems they are attacking, the universe of proven programs they are expected to draw on is too cramped.

## TOWARD AGREEMENT ON FUNDAMENTALS

**E**xperimentalists and Inclusionists agree that rigorous evidence should be the basis for decision making in the social policy sphere. Differences arise over what evidence should be considered "rigorous." Current efforts to generate and use many different kinds of evidence suggest that it is possible to attenuate the divisions between the camps. Experimentalists and Inclusionists could probably come to agreement around four fundamental principles: Begin with a results framework; match evaluation methods to their purposes; draw on credible evidence from multiple sources; and identify the core components of successful interventions.

**Begin with a results framework** | Experimentalists and Inclusionists can agree that an essential first step for those embarking on an initiative is to identify the clear, measurable results for children, families, and communities sought by the intervention. Using a results framework from the beginning provides clarity of purpose, helps build a commitment to data, accountability, and performance, and provides a structure for evaluation.

For example, cities participating in the Annie E. Casey Foundation's Making Connections initiative brought together representatives of schools, health and human service agencies, workforce and employment programs, city and county governments, neighborhood residents, United Ways, and local philanthropies to agree on results and to track whether and how new policies, programs, and practices were having the desired effects. The results focus turned out to be a formidable force for change, informing decisions to expand strategies or change course, and bringing attention to what it would take to achieve ambitious, long-term goals.<sup>9</sup>

To facilitate work within a results framework, collaboration to identify and develop compatible outcome measures for interventions aimed at similar problems is essential. This work cannot be done efficiently at the local level. Although sites can use data from schools, municipal services, and health and social service agencies, and can undertake special purpose surveys of residents or service providers to generate additional data, the tasks of securing data-sharing agreements, organizing and managing the data, and collecting new data all require significant resources. Because measures and definitions are rarely standardized (what constitutes child abuse, for example, varies widely from place to place), local indicators are difficult to compare across sites or against national statistics.

Coherent indicators and measures could be developed for specified target populations or for specific measurement challenges. An imaginative example of the latter is the Benchmarking Project, a collaboration of the Annie E. Casey Foundation and Public/Private Ventures. The Benchmarking Project is now helping 159 organizations in three cities to move toward more consistent definitions of performance measures, to adopt or adapt technology, to facilitate exchange of information, to provide more useful reports about local and state data trends, and to learn from research and from peers.<sup>10</sup>

**Match evaluation methods to their purposes** | RCTs are often seen as the gold standard of evaluation, but RCTs are not always the best method for obtaining needed knowledge. The key is to match evaluation methods to specific types of interventions and different needs to know.

Randomized experiments are most useful when we assess interventions that are neatly circumscribed, with a clear causal relationship to the outcome. That is why they work well in medical research. Early education is a harder domain in which to obtain quantitative results, but it too provides examples of useful trials. Take North Carolina's Abecedarian Project, designed in the 1970s to provide intensive high-quality childhood education from infancy to kindergarten. Following participants into adulthood, researchers found that they had better scores on math and reading tests while in school, and less teenage parenthood; they were four times more likely to earn a college degree than the control group. Along with the better-known Perry Preschool Project, Abecedarian was the basis for the highly influential calculations by Nobel Prize-winning economist James Heckman showing the substantial payoff from early intervention.<sup>11</sup>

Randomized experiments can also be used to establish the effectiveness of *components* of more complex interventions. Consider Multisystemic Therapy (MST), an intensive family- and community-based treatment program that has been shown "in rigorous, scientific, gold standard tests to be superior to other interventions for adolescents exhibiting severe antisocial and criminal behavior." Taking a more comprehensive approach to working with troubled youth that is still evolving, Youth Villages has incorporated the proven MST in a quarter of its in-home services sites in four states.<sup>12</sup>

Non-experimental evaluation methods, on the other hand, help us learn about the effectiveness of interventions that are complex, place-based, evolving, and aimed at populations rather than individuals, and that include too many variables and too few units (for example, communities) to make randomization a reasonable choice. These non-experimental methods are most appropriate for learning from interventions aimed at "adaptive problems," problems that are complex and require interventions that cannot be neatly defined and that involve multiple stakeholders.<sup>13</sup>

The Harlem Children's Zone (HCZ), for example, has developed a network of interventions in health, education, family support, and community building. HCZ's infrastructure seeks to ensure that all the children in a designated area (the zone) stay on track from birth to college graduation and entry into the job market. Because this pipeline of supports and interventions is hard to reduce to quantifiable metrics, and because HCZ's mandate is to serve all of the children living in the zone, HCZ as a whole, unlike its individual components, cannot be evaluated with experimental designs.

When we use non-experimental methods, 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 policies under consideration, and not from factors, such as selection bias, that produce markedly non-comparable populations.

Chicago's New Communities Program (NCP), for example, is learning from the workings of its community initiatives without thrusting them into an experimental straitjacket. Evaluators examine demographic changes as well as the nature, extent, and pace of progress in such neighborhood indicators as crime rates, housing market activity, and commercial vitality. Their analysis shows in real time how trajectories vary across NCP communities and how they compare to changes in selected non-NCP neighborhoods and in Chicago overall.

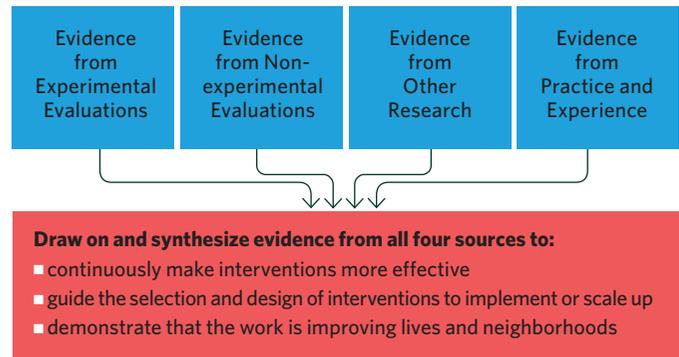
**Draw on credible evidence from multiple sources** | The information for designing more effective interventions and for guiding scale-up and implementation efforts resides in a variety of sources, including program evaluations, other kinds of research, and practice. (See "An Inclusive Evidence Base" at right.)

One example of an effective intervention built on evidence from multiple sources is the sharp reduction in central line-associated bloodstream infections (CLABSIs). Because the remedies that make up this intervention were hypothesized so soon after the urgency of the problem was established (about 43,000 CLABSIs occurred in hospitals in 2009 and nearly one of five infected patients died as a result), the intervention was implemented in the absence of experimental evidence. Each of the core components of the intervention, however, came from some kind of reliable evidence, including evidence-based protocols, the training experience of each of the hospitals, support from statewide, regional, and national collaboratives, multidisciplinary teams that ensured compliance, and the provision of performance data to staff. The effectiveness of the intervention was documented simply by measuring the outcomes as they occurred before and after its implementation. Although CLABSIs are one of the most common types of medical errors that occur in hospitals, the intensive care units of the four hospitals that implemented the intervention brought their central line infections to zero.<sup>14</sup>

The Nurse-Family Partnership (NFP) provides an example of how a proven program could be made more effective by drawing on a broad knowledge base. NFP, which fields nurses to make home visits to low-income teenagers pregnant with their first child, tested its original model in three RCTs beginning 34 years ago with good, if modest, results, especially in prolonging the interval between first and second births. NFP is widely considered the most rigorously proven early childhood intervention.

NFP's model, however, has been basically frozen in time. To maintain its "proven" status, it has not been adapted to take account of the explosion of knowledge in the last two decades. If communities could build on what we now know from sources beyond program evaluations, they would, for example, draw on the findings of the Harvard Center on the Developing Child that parents involved with substance abuse and postpartum depression were the two most common precipitants of toxic stress in children, and would enlist partners that have worked successfully with such parents. They would draw

## AN INCLUSIVE EVIDENCE BASE



on other evidence to add capacity to deal with domestic violence or homelessness. And they would draw on research documenting that family, friend, and neighbor caregivers care for 41 percent of low-income children under age 5 with employed mothers, leading them to extend home visiting to these critical providers of child care.

**Identify the core components of successful interventions** | The core components of effective interventions are often a better guide to action than are model programs. These core components may be programmatic, or may involve implementation or contextual conditions.

Core programmatic components can be identified by analyzing successful programs aimed at similar goals and extracting their critical common ingredients. This information can be applied to the design of new programs and can strengthen existing programs when they adopt more of the elements that successful programs share.

In a study of 548 programs aimed at reducing recidivism among delinquent youth, the Center for Juvenile Justice Reform at Georgetown University looked for attributes that would extend "evidence-based practice" beyond brand-name models. It turned out that much of the programs' effectiveness could be accounted for by a small number of straightforward factors: targeting high-risk cases and taking a therapeutic approach to changing behavior rather than a control or deterrence philosophy. The researchers concluded that close attention to these core components in the design, selection, and implementation of delinquency programs could provide reasonable assurance that programs with those effectiveness factors would reduce recidivism.<sup>15</sup>

It is tempting to think about core components mainly in programmatic terms, but effectiveness depends as much on the quality of implementation. Kristin Moore, senior scholar at Child Trends, and her colleagues examined experimental, quasi-experimental, and non-experimental research, as well as provider wisdom, to identify implementation features that enhance effectiveness in programs for children and youth. They found the following to be crucial to effective implementation: staff training specific to the program and participant age group; dosage and duration precisely adjusted to the target population's needs; lower participant-to-staff ratios; and interpersonal, individualized approaches to teaching and communicating.<sup>16</sup>

Probably the greatest failures in implementation have resulted from not taking sufficient account of how contextual conditions promote success or failure. As a National Academies report explains, "poor implementation of a beneficial program can come from

unsupportive policies or regulations or a lack of funding...; a lack of organizational commitment, capacity, or leadership; ... or a lack of involvement in or ownership in the program by the community.”<sup>17</sup>

Focusing on spreading the identified components of effective interventions is often more promising than attempting to replicate entire programs, because even proven models are seldom so strong that the program will be successful regardless of the circumstances in which it is replicated. The Carnegie Foundation for the Advancement of Teaching has concluded that “integrity of implementation” is a better goal than “fidelity of implementation” because the former can remain true to “essential empirically warranted ideas while being responsive to varied conditions and contexts.”<sup>18</sup>

## FINDING COMMON GROUND

**T**he time is ripe for all of us to engage in a search for common ground. Berwick states that the new scholarship of improvement requires that we “use all of our senses to acquire, not just data, but wisdom, and not just about parts, but about the whole.” He accurately perceives that we work “in a world of true complexity, strong social influences, tight dependence on local context, a world less of proof than of navigation, less of final conclusions than of continual learning, a world not of certainty about the past, but of uncertain predictions and tentative plans about the future.”<sup>19</sup> In this world we oversimplify at the risk of becoming stuck in the past.

So how do we become unstuck? How do the generators and consumers of information about “what works” come to value both the certainty of findings that come out of randomized experiments and the breadth and depth of the probabilistic findings that come from other research and practice? Strong leadership by public and philanthropic funders could bring this about. Funders can withstand the accusations of being “unscientific” or “soft” when they urge incorporating the concerns of the Inclusionists. Funders can withstand the accusations of trampling on innovation and wanting unattainable guarantees when they urge incorporating the concerns of the Experimentalists. And funders are in a position to reassure the evaluation industry and academics that a carefully designed, inclusive approach to evidence will not sully their reputation for scientific objectivity, nor lessen the demand for their services, but will bring more intellectual challenges.

Funders could support the broadened collection and synthesis of many kinds of evidence from a full range of intervention efforts. They could remind each other that it takes time to produce results, that not every worthy intervention can provide proof of effectiveness, and that very few can provide evidence of significant short-term results achieved on their own. Funders could collaborate with each other and with evaluators to identify and develop coherent measures to compensate for the scarcity of neighborhood- and community-level indicators and to make outcome data more readily comparable. They could encourage more policy-relevant evaluations, help to clarify the circumstances in which experimental methods are and are not useful tools, and refine and legitimize the appropriate use of a range of assessment methods, both experimental and non-experimental.

The broader evidence base and new tools that would emerge from the combined efforts of funders and leaders of the Experimentalists

and Inclusionists could quell some of the current cynicism about what can or can't be done.

To search for remedies is not only compatible with science, but also, as historian Arthur Schlesinger liked to point out, at the core of democratic politics. A belief in remedy is the antidote to social indifference and to despair about our capacity to act in common through government.<sup>20</sup> A belief in remedy and problem solving would have the Experimentalists and the Inclusionists reconcile their competing approaches to evidence and provide reformers in many domains the support of their combined best insights. This would enable hardworking practitioners throughout the country to achieve the breakthroughs so urgently needed by the children, families, and communities now at the margins of American society. ■

Parts of this article are adapted from an article by Lisbeth B. Schorr and Frank Farrow, “Expanding the Evidence Universe: Doing Better by Knowing More,” Center for the Study of Social Policy, July 2011.

## Notes

- 1 Steven Goldberg, *Billions of Drops in Millions of Buckets: Why Philanthropy Doesn't Advance Social Progress* (Hoboken, NJ: John Wiley & Sons, 2009).
- 2 Jeffrey Zients, “Use of Evidence and Evaluation in the 2014 Budget,” Office of Management and Budget, May 18, 2012. <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-14.pdf>
- 3 The Frontiers of Innovation, *Minds Wide Open: An Action Strategy and Status Report on the Frontiers of Innovation in Early Childhood Policy and Practice*, Harvard University Center on the Developing Child, May 2012.
- 4 Donald Berwick, “The Science of Improvement,” *The Journal of the American Medical Association* 299(10), 2008: 1,182-84.
- 5 Peter Orszag, “Increased Emphasis on Program Evaluations,” OMB Memorandum, October 7, 2009.
- 6 Robinson Hollister and Jennifer Hill, *Problems in the Evaluation of Community-Wide Initiatives*, Russell Sage Foundation, 1995.
- 7 Donald Berwick, “Broadening the View of Evidence-Based Medicine,” *Quality & Safety in Health Care* 14, June 2005: 315.
- 8 Peter Orszag, “Building Rigorous Evidence to Drive Policy,” Office of Management and Budget, June 8, 2009.
- 9 Frank Farrow, “Response Essay,” *Voices from the Field III*, Aspen Institute, 2010; and Leila Fiester and Ralph Smith, “Learning While Doing the Three Rs: Roles, Results, and Relationships,” Making Connections Working Paper, Annie E. Casey Foundation, August 2007.
- 10 Marty Miles, Sheila Maguire, Stacy Woodruff-Bolte, and Carol Clymer, *Putting Data to Work: Interim Recommendations from the Benchmarking Project*, Public/Private Ventures, November 2010.
- 11 James Heckman, “The Economics of Inequality: The Value of Early Childhood Education,” *American Educator*, Spring 2011.
- 12 The Washington State Institute for Public Policy estimated that using MST rather than traditional services saves from \$31,000 to \$130,000 per participant.
- 13 Ronald Heifetz, John Kania, and Mark Kramer, “Leading Boldly,” *Stanford Social Innovation Review*, Winter 2004.
- 14 Sharon Silow-Carroll and Jennifer Edwards, *Eliminating Central Line Infections and Spreading Success at High-Performing Hospitals*, The Commonwealth Fund, December 2, 2011.
- 15 Mark Lipsey, James Howell, Marion Kelly, Gabrielle Chapman, and Darin Carver, *Improving the Effectiveness of Juvenile Justice Programs: A New Perspective on Evidence-Based Practice*, Georgetown University Center for Juvenile Justice Reform, December 2010.
- 16 Kristin Moore et al., “Program Implementation: What Do We Know?” *Child Trends*, 2006.
- 17 Mary Ellen O’Connell, Thomas Boat, and Kenneth E. Warner (eds.), *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities* (Washington, DC: The National Academies Press, 2009).
- 18 Paul LeMahieu, “What We Need in Education Is More Integrity (and Less Fidelity) of Implementation, R&D *Ruminations*, Oct. 11, 2011. <http://rd.carnegiefoundation.org/what-we-are-learning/2011/what-we-need-in-education-is-more-integrity-and-less-fidelity-of-implementation>
- 19 Donald Berwick, “Eating Soup with a Fork,” Nineteenth Annual National Forum on Quality Improvement in Health Care, December 11, 2007.
- 20 E.J. Dionne Jr., “A Historian Who Saw Beyond the Past,” *The Washington Post*, March 2, 2007.