

From the issue dated August 20, 2009

To Judge What Will Best Help Society's Neediest, Let's Use a Broad Array of Evaluation Techniques

By Lisbeth B. Schorr

Having lived in Washington since John F. Kennedy was elected president, I can name the years when the people in the nation's capital who dreamed of making the world better were energized by a new sense of the possible.

This is one of those years, but now the new era of possibility is also one of accountability. Resources are painfully finite, and even the activists and optimists realize that many policy makers — and much of the public — are profoundly skeptical of organized efforts to help those who begin life with the odds against them.

Unless reformers can cite evidence that what they are doing or proposing will indeed make the world better, they won't get support. Not from any level of government, not from increasingly sophisticated philanthropists, and not from the public.

The pressure to support only those programs that have been proven — often referred to as "evidence-based" — seems eminently sensible on the surface. But what does "evidence-based" mean?

Answers to that question provoke intense speculation, cheering, and dismay in blogs and water-cooler conversations. Whether the issue is reconstituting a failing school, reducing child abuse and neglect, supporting isolated families, or starting a new program to visit the homes of people who are ill or facing other troubles, community organizations and their donors want to know what, exactly, merits the "evidence-based" imprimatur.

How evidence is defined will determine whether the demands for evidence will strengthen or undermine the nation's capacity to respond effectively to social needs.

The definition most aggressively promoted today holds that approaches to solving social problems should be considered evidence-based only when they have been found effective by research methods involving random assignment of participants

to experimental and control groups. This narrow definition is an understandable reaction to the era of letting a thousand flowers bloom and allowing good intentions to substitute for sound reasoning about how activities will lead to results. But it is an overreaction.

We have reached the point that the late MIT organizational theorist Donald Schon described as "epistemological nihilism in public affairs," the view that nothing can be known because the certainty we demand is unattainable. And we have done so at a time when richer, more inclusive ways of determining what works are available.

The prevailing definition of evidence is so narrow that its continued ascendancy will inevitably reduce the chances of expanding promising strategies and developing effective new responses to urgent social needs.

Unless we stop ranking possible solutions to problems by their evaluation methodology and find ways to judge how well they accomplish important goals, we will be left with a seriously impoverished tool kit. If government agencies and private grant makers, afraid of being considered not rigorous, unscientific, or wasteful, choose to support only those efforts that meet the randomized-trial test, we will be robbed of:

- Good programs that do not lend themselves to random-assignment evaluations.
- Reforms that are deeper and wider than individual programs.
- Innovations of all kinds.

We risk losing programs that do not lend themselves to random-assignment evaluations because such programs feature multiple interactive components and significant front-line flexibility. And they work best when they can be tailored to unique and changing local conditions and can emphasize hard-to-measure ingredients like respectful trusting relationships.

In contrast, programs that can be evaluated with randomized trials are standardized, remain constant over time and from one site to another, consist of clearly separable components, focus on readily measured efforts, and usually attempt to solve one circumscribed problem through one circumscribed remedy.

Yet we know that single, isolated remedies rarely improve results among the most vulnerable precisely because they change only one thing at a time.

It is the very nature of the most promising responses to persistent social problems that makes them almost impossible to evaluate by the methodologically elegant ways in which we evaluate drugs or electric toothbrushes.

It's not surprising that the registries of programs with proven results are woefully thin.

For example, the Coalition for Evidence-Based Policy, established to correct what it calls the "central problem" — that America's social programs "are often implemented with little regard to rigorous evidence" — labels only three "early childhood interventions" as effective.

Surely we want communities with new funds to invest in early childhood to have more to choose from than one program from the 1960s that served 123 children, one university-sponsored program from the 1970s, and one — currently in operation — providing nurse home visits to first-time teenage mothers.

Unfortunately, the narrow definition of what qualifies as evidence-based (and therefore worth investing in) has caused sponsors of many promising but complex solutions to social problems to retreat.

I have been at meetings where knowledgeable, experienced program managers argued to adapt a model program to meet varying circumstances in each site but were persuaded by evaluators to change their plans in the interests of "evaluability" by randomized trials. They were pressured into trading the probability of effectiveness for the certainty promised by a randomized evaluation.

The essential elements of significant change in the world beyond programs — the infrastructure to monitor changing community needs and facilitate connections across agencies, systems, silos, and financing streams — can be built on lessons from experience but not on models proven effective by experimental evaluations.

We already see government agencies and private foundations shrink from approaches that don't fit into a programmatic box — things like efforts to get more-talented, better-trained, and better-compensated staff members into schools and child-care programs that serve the most disadvantaged children, or to mobilize mental-health professionals to consult with home visitors or classroom teachers, or to build greater trust between communities and helping institutions.

Government agencies and foundations may recognize that much of what needs doing is not amenable to programmatic solutions. But they don't act on that knowledge because the evidence of effectiveness that evaluators and policy makers seek is so hard to extract and takes so long to appear in changes that reach beyond single programs.

Many of our social problems require new solutions. When support goes only to what has been shown to work in the past, however, the impetus to find new responses shrivels and we remain mired in the status quo. By definition, "evidence-based" is about what worked in previous decades, not what is likely to work in the next.

Even the field of medicine, which gave randomized clinical trials their heyday, has outgrown their constraints. The Roundtable on Evidence-Based Medicine of the Institute of Medicine called for a re-examination of what constitutes evidence and suggested that randomized clinical trials should not be considered the gold standard.

The roundtable declared that medicine's dependence on the trials is inadequate today and may be irrelevant tomorrow, as it seems to be useful only in increasingly limited circumstances (including a narrow range of illnesses and the absence of multiple problems in an individual patient).

It's hard to break with the dogma of experimental design as the sole source of reliable knowledge. Policy makers cling to experimental design because they want the proof that comes from incontrovertible numbers — and, for some, if that results in fewer programs to support so much the better.

A more inclusive approach to building knowledge would offer the wide range of useful data that we need to improve the prospects of vulnerable children and families on a large scale.

A more inclusive approach applies experimental techniques, including randomized trials whenever appropriate, but only when appropriate. It makes use of multiple methods, to draw inferences from multiple sources of evidence analyzed in the context of sturdy theory.

By using a more inclusive approach to judge what works, organizations could move away from oversimplified success-or-failure judgments to a richer knowledge base about approaches that are plausible, promising, or proven. People designing social programs would have at their fingertips the lessons learned from theory, research, and experience, enabling them to construct ever stronger hypotheses, and ever more effective ways to solve problems.

So if we risk losing too much with a too narrow definition of what constitutes credible evidence, how do we minimize the risks of squandering money and effort — and people's hopes — on efforts that do no good?

Can we find ways to select, finance, and design solutions, judge their progress, and hold them accountable that don't rely on anecdotes, good intentions, or ideology or on rigid methodologies that don't fit the interventions they are assessing?

Yes, we can.

We can develop approaches that are rigorous and informed by the evidence even when they don't involve experiments.

One attractive alternative is to adopt a "results framework." Sponsors at the national, state, or local level select outcomes that the public values (e.g., more babies born healthy, lower rates of child abuse and neglect, more children reading proficiently at third grade) and the indicators that measure progress toward those results.

Then the community organizations determine — on the basis of research, theory, and experience — the actions likely to contribute to attaining those goals, be they proven or promising approaches, new combinations of programs, stronger infrastructure, new capacities, or the development of innovative efforts.

For example, the probability that children enter school ready to learn increases with the availability of high-quality prenatal care, good health care once the children are born, child care that fosters learning, and early identification of and responses to developmental problems. There is strong empirical evidence, for example, that social isolation is a major risk factor for poor outcomes, including child abuse and neglect, and children failing at school.

Communities that set out to reduce the incidence of child abuse and neglect and increase the incidence of school readiness will quickly find that virtually no programs have been proven to reduce the risk of social isolation. But they can draw on lessons from hard data, theory, observation, research, and practice to find or construct programs and practices whose success is probable.

For example, an entrepreneurial community organization could collaborate with a successful home-visiting program, a mental-health consultant who helps child-care workers identify struggling families, and a group that deals effectively with maternal depression and drug abuse.

Together, the partners may devise new approaches to dealing with social isolation that we wouldn't want to squelch merely because no one has yet proven that this combination of potential solutions is effective. From such innovations we can identify the contributions to valued outcomes even when we can't make causal attributions.

The issue of how we define evidence has never been timelier. The Obama administration is about to embark on an exciting new Promise Neighborhoods program, inspired by the Harlem Children's Zone.

Applicants for planning support will have no chance to meet this federal effort's ambitious goals if they rely only on programs that fit the narrow definition of "evidence-based."

Harlem Children's Zone itself has been described as an endeavor that "meshes educational and social services into an interlocking web, and then it drops that web over an entire neighborhood."

We won't find interlocking webs or web drops in the directories of evidence-based programs, now or ever. Nor is the problem solved by evaluating the impact of each discrete program, because the entire point of efforts like the Harlem Children's Zone is that we expect the whole to have a far greater impact than the sum of its parts.

We cannot hide from the fact that, as social problems have become deeper and more complex, so have effective ways to solve them, and that makes them harder to assess. Highly circumscribed, standardized approaches that can be assessed by random-assignment studies should be evaluated that way.

But the chances of achieving meaningful social change in today's world are sharply reduced if we fail to recognize that this methodology is only useful in a small proportion of real-world circumstances. Unless we embrace the alternative approaches that incorporate many ways of knowing, many sources of knowledge, and more-inclusive methodologies, we will be robbed of essential information, and the nation's children and youths will be robbed of a more hopeful future.

Lisbeth B. Schorr is a senior fellow at the Center for the Study of Social Policy and a lecturer in social medicine at Harvard University. She is also the author of Within Our Reach and Common Purpose.