Malcolm Wiener Center for Social Policy

Working Paper Series

www.hks.harvard.edu/socpol/publications_main.html

A Lot to Lose:

A Call to Rethink What Constitutes "Evidence" in Finding Social Interventions that Work

By Katya Fels Smyth and Lisbeth B. Schorr

January, 2009

A Lot to Lose:
A Call to Rethink What Constitutes "Evidence" in Finding Social Interventions That Work

by Katya Fels Smyth and Lisbeth B. Schorr

A growing emphasis on accountability has led policy makers, funders, practitioners and researchers to demand greater evidence that program models "work" and that public and private dollars invested are generating relevant results that can be directly attributed to the given intervention. The gold standard for making these judgments is presumed to be the experimental–design study. In this paper, the authors suggest that the underlying assumption that everything that "works" can be judged with the same methodology has dramatic negative consequences for the field, for funders, and for those that desperately need high quality programs. The authors describe the characteristics of *What It Takes* organizations, which their work suggests support lasting change in the lives of highly marginalized and vulnerable people. They describe the ways that experimental methodology is a poor fit for judging the impact of these program models, while they find insufficient use of more appropriate ways of assessing their impact. They identify the risks inherent in the continued privileging of experimental designs over all others, and suggest that the risks are heightened in periods of great economic stress, when the pressure for accountability is increased. The authors suggest a set of starting points for rethinking evaluation to ensure greater accountability without reducing the chances that those who need help the most will have access to programs that support meaningful, lasting change.

Contact information:

Katya Fels Smyth
Research Fellow, Malcolm Wiener Center for Social Policy, Harvard Kennedy School of Government
Mail: The Full Frame Initiative, PO Box 390955, Cambridge, MA 02139
t. 617-620-6718
katya@fullframeinitiative.org

Lisbeth (Lee) Schorr
Director, Project on Effective Interventions
Senior Fellow, Center for the Study of Social Policy
Mail: 3113 Woodley Rd. NW, Washington DC 20008
t. 202-462-3071
lisbeth_schorr@hms.harvard.edu

Policymakers increasingly prefer to support human service programs that have been "proven" with scientific rigor over those that have not been or cannot be so tested. This preference makes sense on its face. But, paradoxically, the quest for irrefutable proof that a social program or practice is effective may dramatically limit the range of interventions that would solve urgent social problems.

In today's atmosphere of economic crisis, it is tempting to adopt a bunker mentality to protect what we have built to assist vulnerable populations, even if only months ago we were calling for the reform of these same programs, systems or approaches. Should the opportunity arise to add something new (as it may under a new administration) the pressures are intense to go forward with only the programs and models amenable to experimental proof.

In this paper, we urge caution. In assessing the success of efforts to improve outcomes for vulnerable populations, experimental methods must not be the sole arbiter of effectiveness. We pay too high a price when we give credence only to evidence that provides absolute assurance of change in a particular domain, for that threatens to skew our understanding of what constitutes a good intervention that changes lives, not one piece of a life.

We examine how experimental methods are an especially poor fit with the efforts that could help the most vulnerable populations. People who face barriers that interact and occur in clusters must be seen in their real-world contexts, taking into account their challenges and strengths, their relationships and communities. Only then are we likely to be able to respond effectively. Our evaluation methods must be modified to embrace this complexity, not simply to control for it as nuisance variables.

The drive to limit investment of public and philanthropic dollars to interventions of proven effectiveness means that some of the programs that work

best for people and communities facing multiple, daunting challenges will become less able to compete for funds and recognition. This puts the survival of valuable programs at risk and may curb the creation of new interventions. As a result people who are struggling will be deprived of vital, high-quality help, and society will be able to mobilize only a fraction of the interventions that could help those who face multiple challenges to lead productive, decent lives.

Consider the situation of a woman who needs and seeks society's help and support, but for whom prevailing arrangements to help are failing.

> Vanessa regularly brings her children to the emergency room with asthma attacks triggered by damp and mold that seep in through the walls of her converted basement apartment. Her minimum-wage job is threatened because she misses work to care for the sick children, and the school is complaining about their frequent absences. She can't afford better housing and fears that if she reported the landlord's code violations he would evict her—especially since she's a month behind on the rent. She worries that eviction might be the final straw that would prompt child protective services—already involved in her life--to take her children away. She keeps the phone number of the man who fathered her children and often thinks of calling him. If he moved back in, there would be a little more money—but also, in all likelihood, a return to his beatings.
>
> An unusual doctor at the emergency room looks beyond the asthmatic children and recognizes Vanessa's depression. He refers her to a drop-in mental health clinic in the same hospital, where she gets an anti-depressant and a therapy session. The therapist is emphathetic but unable to do much more than to listen and to give her the housing authority's phone number. On a day when both children are well enough to attend school and she has money for bus fare and when she is off work, Vanessa goes to the housing authority. She and her children are placed on a waiting list (years long) for subsidized housing.
>
> Later, leaving her children in a neighbor's care (in exchange for food stamps), Vanessa manages to attend the first session of an evening job-training program. She hoped to find a job that pays more than minimum wage, but she learns that such a switch would disqualify her for the subsidized housing she needs to keep her children dry and healthy, and the program is asking a lot of questions about her current housing situation that make her really nervous about their motives. She leaves. When she gets home she tucks her children into the bed of couch cushions they sleep on and washes down a

> double dose of anti-depressants with some vodka. None of these — not the therapist's empathy, not the anti-depressants, not the possiblity of housing down the road or a job training program -- reaches the core of her, and she is left feeling that she and her situation are hopeless. Every day she interfaces with caring service providers. Every day, she is more and more lonely.

Although Vanessa's challenges can be defined as discrete issues (poverty, depression, lack of job skills, children's illness, housing instability, etc.), they connect in complex ways. Her life isn't a braid that can be separated into distinct threads and sorted into a logical pattern—her life, her issues and strengths and context are, like they are for all of us, co-mingled. Isolated responses, therefore, don't offer much hope. The siloing of services may help providers rationalize the mess, but it often diminishes the services' power and undermines needed supports, paralyzing those whose lives are messy. Efforts to integrate and coordinate services (often through "one-stop shopping" centers that house multiple providers) also fall short, in part because they don't view people's problems as being interconnected (as opposed to simply co-occuring). As such, a host of proven interventions may not add up to a proven whole.

Perhaps more importantly, even most coordinated services fail to recognize and respond to the person and the context beneath the cluster of issues. The drive to focus on and refine the technical aspects of assistance and care have sidelined the equally important elements of responding to the suffering and struggles that add to more than a series of discrete problems. Problems and challenges are not separate from people, but they do not wholly define people, either.

Thus the woman in our example, her children, and millions of others thrash or float through interventions without significant, lasting impact because they fail to engage the core of people's lives — the chronic obstacles that bind one crisis to the next, the extreme experiences (including violence, trauma, poverty, hunger, and illness) that have become customary, the human

relationships that may be as toxic as they are supportive, the unique context in which each person struggles to survive.

The good news is that there are programs that have found ways to help vulnerable and marginalized people, families, and communities make and sustain progress in multiple realms (including health, safety, economic stability, and family cohesion). These programs view people through an ecological lens that encompasses challenges, strengths, relationships, and community context, and they work to craft a response that are "of a piece" with people's lives. We call these *What It Takes* programs.

The bad news is that *What It Takes* programs are increasingly difficult to establish and sustain because of the pressures from funders and policymakers for "scientific" proof of effectiveness. This is not an idle concern, as indicated by the struggles of a small program in King County, WA, recently described in the *Stanford Social Innovation Review.* The program's clients (mostly recent immigrants) struggle to survive without adequate food or clothing, assaulted by mold and gas leaks in an apartment complex built 70 years ago as temporary housing for war veterans. The program's director and evaluation consultant described their ultimately futile efforts to determine whether their program worked using experimental methods.

> "Although quantifying the outcomes of flexible, innovative, and holistic programs like ours is difficult, we have tracked our progress for a decade. But now we face mounting pressure to prove, with scientific precision, that our programs positively affect the lives of children and families. Nationwide, a movement to allocate public funds only to evidence-based programs is currently under way. Oregon recently passed legislation that restricts funds to proven effective practices. And although the Washington Legislature did not pass a similar bill this past session, we expect the issue to resurface next year.
>
> "To be 'accountable,' programs are supposed to be evidence-based. But organizations like ours (i.e., small, flexible, community-based programs) do not have the resources to generate the evidence that funders and the public want.

> "[The] kind of rigid, narrow accountability that funders are demanding is of questionable validity …[and will force] programs [to] keep doing only what worked yesterday, instead of what works today. Scientific evaluations generally require staff to standardize interventions and deliver them consistently over long periods of time, regardless of individual needs, cultural considerations, or changes in circumstances. In contrast, [our program] aims to be flexible, innovative, and culturally competent. And so the very qualities that staff and families believe make the program effective are the qualities that make measurement difficult." (Silverstein & Maher, 2008, p. 23)

Because the prevailing pressures for scientific certainty and proof do the most damage to interventions with the most promise for families and individuals who have complex needs, we begin by reviewing what makes these programs effective. We then identify ways that the technology being promoted as the gold standard of evaluation—randomized clinical trials (RCTs) and the experimental method in general—are grossly mismatched to the task of evaluating programs with these attributes. We then describe the risks and losses we anticipate if the field continues to favor RCTs and other experimental methods as the sole source of evidence of effectiveness. We conclude with some preliminary thoughts about complementary approaches to assessment, accountability, and evaluation that show promise.

WHAT WE HAVE LEARNED ABOUT WHAT WORKS

Although they are far from the norm, *What It Takes* programs do exist. Through our collective experience,[1] we have independently identified the mutually reinforcing characteristics of interventions that help people who need more than circumscribed "treatments" to make and sustain lasting, positive change in their lives. Their staffs know that bureaucratic behavior doesn't

[1] This list is drawn from our combined experience: The work of Katya Fels Smyth with the Full Frame Initiative, which focuses on spreading a particular kind of *What It Takes* program; and the work that has gone into the Pathways Mapping Initiative and Lisbeth Schorr's books. See www.fullframeinitiative.org; Smyth, K. F., Goodman, L., and Glenn, C., The Full-Frame Approach: A new response to marginalized women left behind by specialized services. *American Journal of Orthopsychiatry* 76(4) 489-502, 2006; www.Pathwaystooutcomes.org; *Within Our Reach* (1988); and *Common Purpose (1997).*

transform human behavior; they expect and are expected to work in ways that seem above and beyond the call of duty. Staff realize that transformation requires relationships with a high level of trust on both sides, and participants know that the staff won't betray or abandon them. *What It Takes* programs are driven by more than intuition and good intentions, however. We have found that they share five defining characteristics and values:

1. ***An emphasis on relationships and trust***

- All work is embedded in enduring, flexible relationships between staff and participants and in respect for the centrality of relationships.

- Trusting, intentional relationships are recognized as conduits for growth, change, and challenge as well as sources of support and empathy.

- The work setting and the selection, training, supervision, and support of staff emphasize the capacity to form and maintain continuing, respectful relationships that appropriately challenge all parties to do more and to do better.

- The program goes beyond specific relationships to generate a sense of community and of belonging to something positive, not something remedial.

2. ***An orientation toward working in partnership with program participants***

- Participants' own narratives, intentions, and concerns are valued aspects of the relationship between participants and staff.

- The default expectation is that staff will do whatever needs to be done rather than adhering to a rigid job description. There isn't a firm boundary, for example, between staff who provide counseling and those who help a person move from one apartment to another.

- Programs are organized to make sense to program participants, even if that means more work for program staff.

3. Significant front-line flexibility within established quality standards

- Programs with multiple sites grant control over intake and recruitment to local staff, within broad parameters.

- Programs grapple with the tension between helping people make measurable progress toward a specific goal (e.g. getting a job, staying sober) with the challenges of maintaining that progress when other factors can't be shifted as easily (e.g., long waits for subsidized housing).

4. A deep understanding of the importance of the larger environment

- Multiple program components respond to both children and adults in family, peer, and neighborhood contexts.

- The program takes context into account. It recognizes the immediate context—a combination of current conditions and personal and cultural history-- as something that creates and shapes interrelated health, social, cultural and educational needs and preferences. It also takes into account the wider context of economic and bureaucratic pressures within which the program and its clients operate.

- The intervention reflects local strengths, needs, and preferences and evolves in response to experience and changing conditions.

- The program helps participants to manage other stresses in order to sustain the progress that comes about when partner organizations with specialized expertise respond to participants' specific challenges, such as a medical problem or a threatened eviction.

- Boundaries between the program and its geographic location are porous, making the program invested in the health and sustainability of community residents who may not benefit directly from its services.

- Programs may engage in advocacy and policy work to cause changes that benefit a group much broader than program participants.

5. Accountability

- Staff and management judiciously use data—quantitative and qualitative—to continuously refine and improve program design and practice, as well as to document the impact of current practice.

- Program staff and managers strive to reflect on, understand, document, and maximize their impact without prescribing specific goals or outcomes for individual program participants.

- A climate of relentlessness drives staff not to give up even if a person, family, or program has to give up on a particular outcome in a particular timeframe. In other words, staff are more accountable for sticking with an individual than for producing a specific outcome.

- The program recognizes that personal change takes hard work, that sustaining change may be even harder, and that helping people sustain change may require different strategies from those employed to help them make initial changes. A *What It Takes* program invests in helping people make changes, and in supporting their efforts over time.

To be sure, these programs are rare in part because it is hard to work in these ways. It takes a certain kind of staff person and a certain kind of support for that staff person to hold everything together. It takes work environments that are able to loosen or dodge bureaucratic constraints. And it takes program administrators who can and will champion this different way of doing business in order to secure funding and public support. But these programs do exist, and they can make the difference between a downward spiral—punctuated, perhaps, by short-lived "successes" achieved by participating in specialized programs alone—and a healthier, more settled life for people and families who have rarely known either.

"PROOF" IS NOT ENOUGH: LIMITATIONS OF THE EXPERIMENTAL METHOD

The last several decades have seen a push to understand what works well and for whom, what works less well, and what doesn't work or works but only at such high cost that it is a poor value. We count ourselves among the legions arguing that managers and staff of interventions aimed at improving lives among troubled children, youth, adults, families, and neighborhoods must be accountable for doing effective work. They owe accountability not only to themselves and to their public and private funders, but also to those who come to them for assistance.

The authors' issue is not, and has never been, with the principle of accountability, but with the limited technologies in use to establish accountability,

which make it so hard for *What It Takes* programs to be understood and assessed. It isn't that these programs cannot demonstrate their value; it is that the ways that they are asked to do so are poorly aligned with what they actually do, the ways they create greatest value, and the outcomes they seek to achieve. There is a fundamental mismatch between the task of understanding the workings and impacts of these programs and the prevailing assessment tools and mindsets. Consequently, *What It Takes* programs often are misunderstood and their role undervalued (except, we have observed, by the people who participate in them).

The technology that provides the current "gold standard" for proving effectiveness is the randomized controlled trial (RCT) used by experimental-design evaluation.[2] Although difficult to construct and costly to administer, the implicit promise of this methodology is that it is pure, decisive and flawless. It can determine definitively whether or not a specific intervention—be it a clinical protocol, a pill, a model program, or a procedure—produces an outcome different than what would occur without the intervention. The evidence produced by RCTs provides incontrovertible numbers, statistical analyses of p-values, and findings of causality that discern order beneath myriad human interactions. Programs and practices that demonstrate results to a level of statistical relevance are branded "proven," and the funders and program designers who select them feel

[2] Gordon Berlin, president of MDRC, describes randomized controlled trials as follows: "To determine the net difference a program makes, one needs a counterfactual, a comparison (or control) group of similar people that shows us what would have happened in the absence of the program. The most reliable way to create a counterfactual or control group is to use a random assignment research design — widely accepted as the "gold standard" — essentially the same research method used in medical research to determine the effectiveness of a new medicine. Random assignment uses a lottery-like process to create two groups that do not differ systematically — except that one is eligible for the new program and one is not. By identifying a pool of eligible people, and then randomly assigning them to a program group that is eligible for the new services or to a control group that is not, any subsequent difference in outcomes between the two groups — say, employment rates — can be confidently attributed to the effects of the program. Random assignment designs are fair: everyone has an equal chance to participate in the program…. The results from random assignment studies have the virtue of being simple to understand, and, when implemented well, such studies are seldom challenged." (MDRC, 2007).

confident that they are supporting or implementing something certain to deliver positive results. The data, not fallible decision-makers, can be said to make the case for “what works.”

An impressive set of model programs and interventions that meet the experimental-design test have emerged, including the Carolina Abecedarian Project, Chicago’s Child-Parent Centers, the Incredible Years, the Infant Health and Development Program, Multisystemic Therapy, and the Nurse Family Partnership. These and other programs that can be proven effective by experimental means share several characteristics. Their elements are circumscribed and clearly defined. At least while they are being tested, they don’t evolve over time or change in response to contextual factors. They are sufficiently independent of the particulars of place, funding structures, and policy contexts that their methods and results could indeed be replicated with fidelity to the original model. (See Figure 1, “Contrasting characteristics of programs that are and are not appropriate for experimental evaluation.”)

For this category of intervention, experimental methods are a useful filter for sorting “what works” from what doesn’t work.

Interventions whose program design will not allow experimental evaluation, meanwhile, are deemed unproven—and, to many funders, “unproven” equates with a passing fad or an idea that is unlikely to deliver concrete results. That is why we see public and private funders increasingly requiring that applicants for support show that the program they are proposing has been found to be “evidence-based.” (See Figure 2, “Examples of rising demands for narrowly defined evidence as a basis for funding.”)

Economist Rob Hollister says, randomized experiments 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." (Hollister & Hill, 2005)

The "flawed alternatives" to randomized experiments may provide less certainty about the causal relationship between intervention and impact, but they do offer a broader range of information that may be more useful in making real-time judgments about the real-world effectiveness of *What It Takes* programs and other complex interventions, and will therefore make possible more informed decision-making. New and more inclusive approaches to knowledge building must now be brought out of the shadows and valued for the rich contributions they can make to understanding and strengthening previously neglected efforts to address complex social problems. These new and more inclusive approaches insist on rigor even in the absence of certainty, and find credible evidence of effectiveness in strong theory; an accumulation of empirical evidence from similar or related efforts; consensus among informed observers based on a combination of theory, research, and practice experience; and a commitment to continually attending to evidence that confirms or threatens an assumption of effectiveness.

Ironically, the field of medicine—the very arena that gave RCTs their original legitimacy—is moving toward a more inclusive approach. The Roundtable on Evidence-Based Medicine of the Institute of Medicine of the National Academies has called for broader approaches to the development of evidence and an end to sole reliance on RCTs. In a 2007 statement, the Roundtable explained that “The prevailing approach to generating clinical evidence is inadequate today and may be irrelevant tomorrow, given the pace and complexity of change. The current dependence on the randomized controlled clinical trial (RCT), as useful as it is under the right circumstances, takes too much time, is too expensive, and is fraught with questions of generalizability.” (Olsen, Aisner, & McGinnis, 2007, p. 5). The Roundtable questioned whether the randomized controlled trial should continue to be considered the gold standard 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 called for a re-examination of what

constitutes evidence and how evidence varies by circumstance, and it suggested that more attention and resources should go to *practice-based evidence* in order to make findings relevant to clinical practice and policy making. (Olsen, Aisner, & McGinnis, 2007).

These developments in medicine and other fields are heartening to those of us who focus on strengthening the services and supports intended to improve the lives of individuals and families who are disconnected from the American dream. It is comforting to know we are not alone in recognizing how severely a sole reliance on experimental evaluation methods has hampered the development of precisely the interventions and supports most important to effectively addressing vulnerable populations and complex problems.

MUCH IS AT STAKE

The risks of continued and increased reliance on narrow approaches to determining "what works" are multiple and serious:

We risk continuing to distort social policy priorities. Interventions that can be assessed by experimental methods attract the bulk of talent and resources, while promising activities that aren't built on a linear relationship between cause and effect and cannot be entirely contained and controlled in a laboratory-like setting will be disparaged and downgraded. Interventions will continue to be ranked by the elegance of their evaluations instead of by their contribution to solving urgent social problems.

We risk not being able to make reliable judgments about the effectiveness of those programs for which experimental-design evaluation is a poor fit. Not all *What It Takes* programs are equally good or equally poor. Indeed, the authors have no knowledge of where on the effectiveness continuum the King County, WA, program described earlier, fits. And we cannot know in the absence of publicly accepted and valued evaluation approaches that can better

illuminate what works in complex settings and for populations with a high number of risk factors.

We risk compromising the effectiveness of *What It Takes* programs that are promising and working. To produce the narrow range of evidence that funders demand, these programs are increasingly pressed to focus their efforts on components that can be studied experimentally over the efforts that may make them effective. As was the case with Vanessa and millions of Americans like her, a constellation of evidence-based interventions may not add up to a proven whole, and yet the nature of evidence-based work is that it compromises the in-the moment ability to respond quickly and flexibly to emergent crises or opportunities.

We risk spending large sums to gather information that arrives too late to inform the most crucial decision-making. Experimental-design evaluations are tremendously costly and take a great deal of time to construct, administer, and analyze. During this period a program may face difficult choices—for example, does it adapt services in response to new needs and risk invalidating three years of data or heed evaluators' pleas to keep the intervention constant but allow critical needs to go unmet? With the economic crisis threatening to push hundreds of thousands, if not millions, more Americans closer to poverty and disenfranchisement, our clinging to solutions that are not highly adaptable to meet rapidly evolving community opportunities and challenges may be all the more wasteful and harmful.

We risk failures of proven models when they are spread without a clear understanding of the critical programmatic or contextual factors that accounted for the success of the original model. If model programs are cloned with the expectation that initial results will predict success elsewhere, the same experimental conditions must apply in multiple settings. While it is reasonable to expect that a laboratory test conducted in Omaha will be replicable

in a similarly equipped laboratory in Ottawa or Oslo, it is harder to conceive that an intervention for suspected child abuse will translate exactly from the South Bronx to Sausalito to San Antonio. It is an even bigger stretch to think that an intervention to support the family described in the opening anecdote would be technically and formulaically the same in each of these cities, given the highly varied policies, benefits, other programs, opportunities, and social contexts. As evaluator Deborah Daro points out, prior success even in multiple RCTs may not predict future success: "A stronger indicator of a program's continued and future success may be its ability to shift its focus in content or service delivery in light of emerging changes in its target population or within the broader social environment." (Daro, 2007, p. 2)

We risk wasting resources on expensive, specialized interventions whose short-term success is significantly undermined by the complications and chaos of highly stressed families and communities. To be valid, an experimental evaluation must carefully limit the number of variables. The study population is often a rarified version of the real world. So when the studied interventions are applied in stressed communities, the success found in the experimental design may be short-lived. A "proven" program may stabilize a chronically homeless person's bipolar disorder, but once he returns to the streets, is robbed of his medication, and finds solace in a nip of vodka, the value of that evidence-based mental health intervention may turn out to be nil.

We risk not examining, understanding, and valuing the role that *What It Takes* programs play in mitigating the undermining effects of poverty and its stressors. *What It Takes* programs seek to provide the essential connective tissue and supports that help people sustain progress made in narrower programs, but the interactions among multiple interventions and their impact is almost impossible to study using traditional experimental design. Broader evaluative techniques are urgently needed to assess how a *What It Takes* program enhances the effects and sustainability of specialized programs.

Without such new tools, we risk mistaking a proven intervention's failure to work under different or more complicated circumstances as a failure of the model, when in fact the disappointing results may be a failure of implementation, of adaptation, or of using the model to leverage or combine with other necessary supports.

The sum of these risks is a greater and graver risk: ***We risk the loss of existing What It Takes Programs and we risk stymieing the creation of new What It Takes Programs.***

BROADENING THE EVALUATION PARADIGM: A MORE INCLUSIVE APPROACH

What would it take to avoid the risks of a continued reliance on narrow approaches to determining "what works?" What would a more inclusive approach to what should be considered valid evidence consist of?

It would include experimental methods, including RCTs, whenever appropriate *but only when appropriate.*

It would include theories of change—the theoretical understanding of the logic that connects actions and resources to results, and that identifies interim milestones showing progress toward selected outcomes, while remaining open to evidence that points to flaws in the logic that undergirds the original theory of change.

It would include such additional methods as case studies, qualitative research methods, comparing cohorts of program participants' progress with their own baselines, with community level data, or with larger data sets.

It would allow for systemic complexity. Rarely is one program both necessary and sufficient to help highly marginalized and vulnerable people, families and communities make and sustain progress. Therefore evaluation tools also have to be able to incorporate not only a program's work, but how that program fits with other interventions. In other words, some of the very factors and situations that the experimental method controls for may need instead to be explicitly folded into an evaluation.

It would privilege adaptation. While organizational leaders are touted for their adaptive capacity, and organizations spend a great deal of time and money developing their organizational adaptive capacity, the current push for evidence-based practice suggests that adaptive leadership is best exercised in service of maintaining a rigid model in the face of change. Evaluation of interventions and programs must be able to privilege mindful, intentional adaptation and evolution.

It would employ a philosophy of "enough." To assure that the perfect does not become the enemy of the good, it would aim to generate enough evidence to make a robust determination of effectiveness (or lack thereof), quickly enough to allow for continuous improvements in program design and implementation, and in resource allocation decisions.

It would include a consensus approach, such as the Mental Mapping process[3] developed by the Pathways Mapping Initiative, to identify the actions that contribute to specified agreed-upon outcomes. Mental Mapping includes the following steps:

[3] For more on the mental mapping process, see http://www.pathwaystooutcomes.org/index.cfm?fuseaction=Page.viewPage&pageId=435

- Convene knowledgeable individuals, including researchers, policy-makers, advocates, and practitioners who are steeped in their respective fields and diverse in their perspectives and beliefs.

- Ask participants to draw on their knowledge and experience to identify the actions most likely to achieve the outcome under consideration, highlighting issues that might otherwise remain hidden and the connections among programs, policies, systems, and institutions.

- In organizing the assembled information, apply reasonable judgments based on a preponderance of evidence from research, theory, and experience.

It would allow programs to draw on what is already known rather than having to independently prove, at great expense and time, the relationships between inputs and outcomes that have already been demonstrated in other realms. For example, public health research increasingly is documenting the significant deleterious health effects of social isolation. Programs that aim to reduce social isolation should be able to document a greater sense of community and stronger social networks and connections to services and informal supports among their clientele and invoke the research of others to inform the improvements in health and social functioning that are likely to follow.

Lastly, a more inclusive approach to evaluation would **take care to define measurable outcomes that matter**. Most front-line programs want a role in identifying the results for which they will be held accountable, and to which they hold themselves accountable. But most cannot, on their own, do the hard work of finding the closest possible fit between the short-terms and long-term results they strive for and the measurable indicators that document progress toward their goals. Front-line programs should be able to count on help from outside in selecting and defining these measures, and the cost of this help must be built

into funding streams, rather than programs having to supply evaluations and proof to funders who are reluctant to pay for evaluation and for all the difficult work to define the questions that an evaluation must answer.

Outcomes that matter may well include whether a program makes a difference in people's sustaining positive change, not just making an initial change. As noted earlier, too often, making change is seen as hard, and sustaining it as easier. We counter that the revolving door of services demonstrates on a systemic level that people may stop drinking for 15 days, find a new job, leave a batterer, begin to take antipsychotic medications, or make a host of other gains, only to return to drinking, unemployment, battering or psychosis. We must acknowledge the importance of sustaining change by funding the activities that help people to maintain progress and by ensuring that evaluations take account of long-term outcomes -- and thereby consider whether a program is contributing to or mitigating the revolving door phenomenon.

PUTTING IT ALL TOGETHER

We are among those who value the push toward greater accountability. This is a time when we need more than ever to know whether what we are paying for actually translates into changed lives. Debates about which results matter, which matters can be measured, and whether it counts if you can't count it are important and worth having. But the discussions have largely been dominated by those who advocate a limited definition of evidence in the pursuit of decisive determinants of efficacy.

In the current economic climate the greater societal aversion to uncertainty may lead to a further entrenchment into very limited and potentially misleading definitions of what works. Times like these demand we revisit what tools we have at our disposal to deal with seemingly intractable social problems, and also demand that we think carefully and deeply about how we understand if

a program is effective. By basing our judgments on many ways of knowing and many sources of evidence, we can avoid the false choice between relying on random assignment experiments versus relying on professions of good intentions, ideology, and a handful of anecdotes. This is precisely the advice of management guru Jim Collins:

> It doesn't really matter whether you can quantify your results. What matters is that you rigorously assemble *evidence*—quantitative or qualitative—to track your progress. If the evidence is primarily qualitative, think like a trial lawyer assembling the combined body of evidence. If the evidence is primarily quantitative, then think of yourself as a laboratory scientist assembling and assessing the data (Collins, 2005, p. 7)

It may well be that private philanthropy, whether from foundations or venture philanthropists, is best positioned to take the lead in breaking with the dogma of experimental design as the one and only source of reliable knowledge. The evaluation industry has too high a stake in the status quo. Social scientists fear they may be considered "unscientific" when they move away from the randomized controlled trial (McCall & Green, 2004). Government agencies increasingly require that grantees use RCTs to demonstrate effectiveness as a shield against exposés for using scarce public funds to support programs that fail. Individual programs that challenge the gold standard risk losing precious dollars in the competition for scarce resources, or being branded "unaccountable."

Our goal here is not to set forth a single evaluative protocol more appropriate for *What It Takes* programs. In fact, we believe that the search for a single protocol may well be fruitless. Rather, we seek to stimulate a conversation about the urgent need to re-conceive what evidence is considered credible and appropriate in understanding and holding accountable efforts to bring about lasting change in America's most challenged communities. A course of

exploration, discovery, and field building must occur, which itself will require significant funding and vision.

We are not so naïve to suppose that a change in what decision makers consider “evidence” will directly reduce addiction, poverty, child abuse, or other social ills. But we do suggest that the current drive for certainty catches vulnerable, marginalized individuals, families, and communities in a crossfire and may limit the creation, growth, and strength of the very programs that offer a promise of better odds. And we do suggest that the stakes are higher than ever.

And so we close where we began this article, with those who are not “making it” in America today, who are embedded in a nexus of failed social polices, inadequate human service systems, and dashed personal and community aspirations. Tonight, Vanessa may find herself again in the emergency room to get her child an evidence-based dosage of asthma medicine. Unless she is lucky enough to visit one of the handful of hospitals where physicians can refer her to a specially trained cadre of staff or volunteers, nothing in his protocol gives him reason or leeway to help her address the mold or in her apartment. She still hasn’t found a program that will provide the job training she needs, help her locate high-quality childcare, and navigate the trade-off between growing income and shrinking housing subsidies. No one is being held accountable for helping her in the way she needs help. And so the question of how to assess effectiveness has real and raw consequences for this family and tens of thousands of real families across the country.

More than ever, profound changes are needed so that those left behind by America’s progress and prosperity can not only achieve but also sustain decent lives. Among these, a broader, more inclusive and ultimately more accountable approach to how we judge “what works” looms large.

Figure 1

Contrasting characteristics of programs that are and are not appropriate for experimental evaluation

INTERVENTIONS NOT APPROPRIATE FOR EXPERIMENTAL EVALUATION (*What It Takes* Programs)	INTERVENTIONS APPROPRIATE FOR EXPERIMENTAL EVALUATION
Emphasis on relationships and trust: Developing and maintaining healthy relationships is considered both a means to help people make and sustain positive change and an end in itself. Program participants' growth in and leverage of relationships is an outcome to be assessed individually. Consequently, long-term engagement is seen as a necessity.	***Emphasis on provision of services;*** To the extent that relationships are developed, they are a means to provide a specific service. Extended engagement with an individual or a family is seen as fostering dependence. The ability to provide a service and then disengage is highly valued.
Orientation to working in partnership with program participants: The program is designed to respond to each participant's unique cluster of issues, although some (such as poverty or a history of trauma) may be very prevalent. The program responds to and engages the participant as someone who is more than simply a jumble of problems. Goal setting is a mutual, dynamic process between participants and staff. The goals of funders and other stakeholders are considered but are not the sole driver of program goals.	***Orientation to problem solving:*** Programs are designed to address a specific challenge or cluster of challenges in a defined population. Programs do not consider or engage significant aspects of participants' experiences that seem separate from the primary problem. Partnership between staff and participants may occur within the intervention process but far less around framing the desired outcomes.
Emphasis on front-line staff flexibility Staff are expected to exercise discretion in tailoring interventions to situations and goals of individuals, and their efforts to do so are supported.	***Emphasis on consistency of method*** Deviation from the protocol or the model to adapt to a participant's situation or to the intervention's larger context is discouraged and devalued. Such adaptations compromise the label of "proven."
Programs adapt to respond to specific community situations and to changes in context and events Interventions are designed by combining local wisdom with theory and understanding about what has worked elsewhere. Context	***Adaptation to context is devalued and compromises evaluation methodology*** Interventions are not redesigned based on contextual factors. Contextual factors are considered primarily in locating collaborative

informs program design, delivery, infrastructure and partnerships.

Accountability is dynamic

Evaluation is on a feedback loop. Learning and course corrections are continuous, and draw on the program's own experience, as well as the experience of others with similar goals.

Accountability to individuals and to specific outcome may diverge

Individual goals may differ, and the need to achieve a specific outcome does not trump the need to stick with an individual participant.

Measures of programmatic success are broad

Programmatic success is not determined by the frequency of program participants who achieve a goal but by the frequency of program participants who make and sustain change in one or more broad realms (e.g., health, safety, stability). This allows for individual tailoring of program process and goals.

partners and obtaining funding, but not in shaping the intervention model itself.

Accountability is not dynamic and is largely tied to the results of the experimental design evaluation

Program models do not adapt during the course of the study so as not to compromise the validity of results.

Accountability to individuals is tied to accountability for producing specific outcomes

Program goals are pre-determined and not significantly variable among program participants.

Measures of programmatic success are narrow

Programmatic success is determined by the number or proportion of participants who make or achieve specific, readily measurable goals.

Figure 2

Examples of rising demands for narrowly defined evidence as a basis for funding

Growing pressures to rigorously assess results of public and philanthropic investments are an important step toward more effective social reforms. However, growing pressures to show that complex efforts to improve the human condition are "evidence-based," are causing alarm among many social reformers -- because the definition of "evidence-based" is often so narrow that the most significant efforts can't qualify. . The narrow definitions of evidence were once primarily confined to academia. But now, philanthropy and government have adopted a similar posture, valuing elegant research methodology and simple, single numbers over more inclusive ways of documenting effectiveness. This stance assumes added weight, of course, when it carries consequences for funding decisions.

The narrowness of "scientifically based research" in the No Child Left Behind legislation penalizes holistic reforms

The No Child Left Behind legislation makes more than one hundred references to "scientifically based research," invoking randomized or experimental studies as the basis for funding any intervention, from choosing how to teach reading to reducing dropout rates. To encourage the spread of "proven" reform models, the Department of Education established the What Works Clearinghouse (WWC, http://ies.ed.gov/ncee/wwc/), which lists programs as having "positive effects" mainly on the basis of their being proven through experimental design evaluation.

The problem, as education researchers Frederick Hess and Jeffrey Henig point out, is that only a small fraction of education reforms – and the least important ones at that – are appropriate for and are subjected to the randomized field trials that entities like the WWC are looking for (Hess & Henig, 2008). Strong research has now produced a broad consensus around the conclusion that teacher quality is what matters most in improving outcomes, especially in high-poverty secondary schools. But none of the interventions that most experts believe would, in combination, put more high-quality teachers in high-need schools (including, for example, incentives to raise standards of schools of education and expand their role in post-training supports, use of private resources, and a focus on hard-to-staff fields) are amenable to the kinds of evaluation that the WWC prizes.

Similarly, school reform efforts that would change entire schools or even school districts find they are penalized by the WWC's methodological elegance rules. One district-wide reform model was said by the WWC to have "no

discernable effect" even though it was found by MDRC to contribute to increased student attendance, more students graduating from high school, reduced drop out rates, and improved scores on state tests of reading and mathematics. The difference in findings was because WWC based its conclusions on one study of three schools after just one year of the program – the only study that met its methodological requirements. It excluded the extensive data based on four or more years of implementation in 12 other schools with over 8000 students which led to MDRC's conclusions. MDRC recognized that it was impossible to find an urban school district of comparable size and demographic characteristics to the experimental district, that could be "randomly assigned" to serve as a control group, and therefore sampled the best available control schools and used other analytical tools to minimize the likelihood that findings could be due to factors extraneous to the reform effort. (Connell, 2008)

It is no wonder, then, that reformers and innovators worry that resources will be shifted from their ambitious and complex long-term institutional change efforts, to new curricula and other more readily defined, piecemeal changes that can be evaluated more traditionally.

What's easy to count squeezes out complexity in OMB's Program Assessments

The Office of Management and Budget's Program Assessment Rating Tool (PART) was instituted by the Bush Administration beginning in 2001 to improve federal agency performance through program rating tools and management score cards. Because OMB threatens to use its ratings to cut off Federal funding, and has used poor ratings to justify funding cutbacks in its budget proposals, PART exerts heavy pressures on agencies to document their accomplishments on the PART measures. (These pressures are proving to be a burden in many agencies even though there is little evidence that Congress takes the ratings seriously in either the appropriations or authorizing process.)

PART has been characterized by Beryl Radin of the American University School of Public Affairs as a "literal, narrowly focused" approach to program evaluation that dismisses the complexity of local issues, programs, and actors. She cites the example of the HUD's Community Development Block Grant program, which gives cities and states discretion to support housing initiatives, job program, public facilities and more. Under PART, the program was deemed "ineffective" because its multiple purposes and broad scope "created ambiguity." (Radin, 2008, p. 251)

One of the earliest Federal programs to fall under PART review was the Child Abuse Prevention and Treatment Grants program of the Department of Health and Human Services, which provides funds to States to improve their child protective service systems. It was found by OMB to have "a clear focus and well-defined mission, be well-targeted, and free of major flaws that would limit its

effectiveness and efficiency." Yet it originally received a NOT PERFORMING rating, based on the program's failure to achieve a single numerical goal set by OMB: the reduction to 7 percent of reports of repeat maltreatment of children in the States' child protection systems in the program's first two years. (The reduction that was documented was from 9% to 8%.) Its current rating is RESULTS NOT DEMONSTRATED. [4]

The assumption by PART that this one measure would accurately reflect the program's effectiveness is another symptom of the programmatic distortions imposed by a highly over-simplified view that what's readily countable can be used to demonstrate true accountability. It flies in the face of widespread agreement that no single indicator can capture the impact of complex efforts that, in this instance – by legislative mandate – include addressing the needs of drug-exposed infants, finding new ways of referring children not at imminent risk to community or preventive services; and bringing about interagency collaborations across child protective services, health, mental health, juvenile justice, education, and other public and private agencies. It's not that programs of this kind can't be assessed; the difficulty is that assessments that would reflect their real impact can't be squeezed into the Procrustean bed of a single, simple measure.

Even philanthropies that can take risks skew funding requests to the readily countable

The pressures for results that are easy to count that dominate public funding also permeate philanthropic grant making. In the last several years we have been hearing ever more frequently from front-line reformers about how they are reluctant to apply for foundation funds that would support work whose payoff will not be demonstrable, at least in the short run, on terms that funders are demanding. They may have discovered that the greatest current need in a given community is to train and field mental health consultants who could provide continuing support to both staff and families in child care centers and family day care homes to improve outcomes for troubled children. They are familiar with evidence from elsewhere that by reducing social isolation, treating maternal depression, and coaching both staff and parents, they are able to strengthen the protective factors that predict improved outcomes. But the cost in dollars and time and human resources that would be required to prove that their particular combination of these interventions actually results in increased child-well being, higher rates of school readiness and school achievement, as well as less later delinquency, would be prohibitive. So they decide instead to propose a program of eye examinations and follow-up to provide glasses to children with vision defects – an intervention whose effects may be less significant, but are much easier to count.

---

[4] http://www.whitehouse.gov/omb/expectmore/detail/10002142.2004.html viewed November 5, 2008

References

Collins, Jim. (2005). *Good to Great in the Social Sectors: A Monograph to Accompany Good to Great.* New York, HarperCollins.

Connell, James P. (2008). "Clearinghouse's Review of Dropout Programs Faulted." *Education Week* 27(24):30-31.

Daro, Deborah. (2007). "Evidence Based Decision Making: Proven Pathway to Improved Outcomes or Dead-End?" unpublished manuscript.

Hess, Frederick M., & Henig, Jeffrey R. (2008). "'Scientific Research' and Policymaking: A Tool, Not a Crutch." *Education Week* 27(24):26,36.

Hollister, Robison G., & Hill, Jennifer. (1995). "Problems in the evaluation of community-wide initiatives" in *New Approaches to Evaluating Community Initiatives Volume 1: Concepts, Methods, and Contexts.* Connell, James P. and Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss, eds. Washington DC, Aspen Institute.

McCall, Robert, & Green, Beth. (2004). "Beyond the Methodological Gold Standards of Behavioral Research: Considerations for Practice and Policy." *Social Policy Report* 18(2):3-12.

MDRC. (2007). "Experimentation and Social Welfare Policymaking in the United States," " a conference spons. *adapted from a presentation of Gordon L. Berlin. given at "Lancement du Grenelle de l'insertion: Les rencontres de l'experimentation sociale, a conference sponsored by the French Government,* (p. (www.mdrc.org/publications/467/presentation.html)). Grenoble, France.

Olsen, LeighAnne, Aisner, Dara, & McGinnis, J. Michael. (2007). *Institute of Medicine Roundtable on Evidence-Based Medicine: The Learning Healthcare System, Workshop Summary.* Washington DC, National Academies Press.

Radin, Beryl. A. (2008). "Performance management and intergovernmental relations" in Conlan, Timothy J. & Paul L. Posner, eds. *Intergovernmental Management for the 21st Century.* Washington DC, Brookings Institution Press.

Schorr, Lisbeth B. (1997). *Common Purpose: Strengthening Families and Neighborhoods to Rebuild America.* New York, Doubleday.

Schorr, Lisbeth B. (1988). *Within Our Reach: Breaking the Cycle of Disadvantage.* New York, Anchor.

Silverstein, Laura, & Maher, Erin J. (2008). "Evaluation blues: How accountability requirements hurt small, innovative programs the most." *Stanford Social Innovation Review* 6(1): 23.

Smyth, Katya F., Goodman, Lisa, & Glenn, Catherine. (2006). "The Full-frame Approach: A New Response to Marginalized Women Left Behind by Specialized Services." *American Journal of Orthopsychiatry* 76(4):489-502.