STANFORD
GRADUATE SCHOOL OF BUSINESS
STANFORD SOCIAL INNOVATION *review*

Drowning in Data

By Alana Conner Snibbe

Stanford Social Innovation Review
Fall 2006

DROWNING in DATA

In the frenzy over accountability, funders, donors, and the general public are calling for more program evaluation. But few understand evaluation well enough to conduct or bankroll high-quality studies. Without sufficient knowledge or funding, nonprofits are often collecting heaps of dubious data, at great cost to themselves and ultimately to the people they serve.

Rubicon Programs Inc. is one of the lucky ones. The San Francisco Bay Area social enterprise, which employs and aids people living in poverty, has a powerful data collection system that streamlines evaluation and reporting to funders. And Rubicon needs it. Although the organization earns over half of its $15.3 million annual budget from income-generating projects – a locally famous bakery and a landscaping service – it has over 100 different funders.

Before Rubicon installed its data system, the organization faced a struggle familiar to many nonprofits, says Rick Aubry, the organization's executive director. Each funder demanded different data, reported on different forms. The result was that many staff spent large chunks of their time generating the one-off data, while several directors spent large chunks of their time filling in the one-off forms.

Despite all of the time and money that Rubicon invested in creating these reports, they contributed little to improving the program's effectiveness. Funders seldom asked Rubicon to explore ways that it could improve its services. Instead, they often wanted to know only how Rubicon spent their money. These reports "added zero value to our decision making, and did not help us improve our services," says Aubry.

With its powerful new evaluation system in place, Rubicon can now deliver data to its myriad funders in all kinds of permutations, with time and resources left over to collect the numbers that it wants for itself. Ironically, the system has uncovered a new problem: Most funders don't actually care about the data.

"Everyone says they want to be data-driven in their decision making. But now we have all of this robust data, and it doesn't seem to have any effect on funders' decisions," says Aubry. "From the viewpoint of financial sustainability, we are

PHOTOGRAPH BY STEVE COLE/MASTERFILE

by ALANA CONNER SNIBBE

no better off than before."

Rubicon's evaluations fall short on another front. For all their sophistication, they cannot prove that they are "making an impact" – a requirement that many funders now demand, though few understand what impact evaluations entail, and even fewer bankroll them. The only surefire way to show that a program is helping people more than they could have helped themselves is to conduct expensive and lengthy studies involving control groups. Because so many people underestimate the difficulty and cost of demonstrating impact, nonprofits often collect reams of data that are not only useless, but also misleading. As a result, evaluation is failing to help make the social sector more effective.

The Winds of Accountability

"Evaluation is often something that funders want to be seen doing, but not what they value being done," observes Thomas D. Cook, a professor of sociology, psychology, education, and social policy at Northwestern University and a world-renowned expert in education evaluation. "They're feeling the winds of accountability, and they're passing it on to their programs."

Those winds are blowing even colder in the slipstream of the Enron, WorldCom, and United Way scandals, resulting in more demands for evaluation. Nonprofits are bearing the brunt of this evaluation frenzy. A survey of more than 300 nonprofit executive directors and CEOs in New York and Ohio reveals that 75 percent feel that they do not have enough time for evaluation, 61 percent feel that they do not have enough staff for evaluation, and 45 percent feel that they do not have enough funding for evaluation.[1]

Despite the extent of the problem, nonprofits' woes are barely on funders' radar. That's because grantees are reluctant to complain about the situation when they're dependent on funders for their survival, finds a study of more than 200 randomly selected San Francisco Bay Area nonprofits.[2]

Despite the ire it has raised, the demand for evaluation shows no signs of slowing. This is probably for the best. Trusted with other people's cash and absolved of taxation, nonprofits and foundations should evaluate themselves and their programs to reassure the paying public that their money is actually making a difference.

But conducting evaluations that are truly useful is much easier said than done. One of the first barriers to good evaluation is the "promiscuity of understandings of what evaluation is," says Cook. The official definition in the *Encyclopedia of Evaluation* (see "Lingo to Go," p. 41) is very broad, leaving plenty of room for squabbling. The most hallowed professional evaluation organizations – such as the American Evaluation Association (AEA), the American Educational Research Association (AERA), and the American Public Health Association (APHA) – "fight like cats and dogs over little things," says Cook, who has presided over the AEA. "It is a source of great disappointment that the evaluation community is not as powerful as it could be because it is not as united as it should be," he says.

Boards and funders don't misuse evaluation because they are dumb or lazy, or even because they are ornery. Instead, their misuses and abuses reflect the fact that good evaluation is extremely difficult.

The issues are many: Should people conduct more summative evaluations – that is, evaluations that judge whether programs "worked" or made an impact – or more formative ones – that is, evaluations that help organizations improve? Should the methods of evaluation be more quantitative or qualitative? Should the evaluators come from within the organization or without? What should be evaluated? Individual programs? Entire organizations? Sets of organizations working toward a common goal?

Academic conflicts and confusions are magnified on the frontlines of the social sector. A first source of puzzlement is all the jargon through which grantees must wade. Add to this confusion the fact that there are no consistent definitions for the jargon, and grantees are positively bewildered.

Even the largest foundations disagree on the meanings of key evaluation terms, reports Victor Kuo, an evaluation officer in the education division of the Bill & Melinda Gates Foundation. For example, "MDRC [an organization created by the Ford Foundation that conducts large-scale evaluations of social programs] defines impact as the difference in outcomes between

ALANA CONNER SNIBBE *is the senior editor of* Stanford Social Innovation Review. *She received her Ph.D. in social psychology from Stanford University, where she conducted a bevy of both quantitative and qualitative studies. Her writings on class, culture, and psychology have appeared in* The New York Times Magazine *and other publications. She is working on a book about human nature.*

LINGO TO GO

The evaluation world is awash in jargon. To help improve communication, *SSIR* has compiled this handy guide to evaluation lingo.

Effectiveness – How well a program produces its intended outcomes in the real world

Efficacy – How well a program produces its intended outcomes under ideal conditions, such as in a laboratory

Evaluation – The systematic assessment of the value, merit, significance, quality, or state of affairs of a program, product, person, policy, proposal, or plan

Formative Evaluation – An evaluation that takes place while a program is ongoing and that provides feedback for improvement

Impacts – Outcomes proven to be caused by a program

Logic Model – A model of how a program will contribute to its specified outcomes

Outcomes – Changes in individuals, organizations, communities, policies, or governments

Outputs – Tangible products that result from a program's activities – such as the number of brochures distributed or the number of people served – that lead to intended outcomes

Process Evaluation – An evaluation of the activities and events that occur as a program is delivered

Summative Evaluation – An evaluation conducted at the end of a program that determines whether the program met its goals

Theory of Change – Assumptions about the nature of a social problem, what its solution is, and how particular actions will lead to the solution

Sources: *Encyclopedia of Evaluation*, ed. S. Mathison (Thousand Oaks, CA: Sage Publications, 2005);
Rossi, P.H. & Freeman, H.E. *Evaluation: A Systematic Approach* (Thousand Oaks, CA: Sage Publications, 1985).

an intervention group and a control group," says Kuo, "while the Kellogg Foundation sometimes uses the term to mean the impact of a program on policy, on community, on the whole system. Those are very different meanings."

Because funders have their own home-brewed definitions, methods, and measures, many nonprofits spend a lot of time tailoring their reporting to each funder's tastes. "Every single grantor we have has a different evaluation tool or format or criteria they want us to use, and we measure all of them," says the executive director of a nonprofit in Northern California who wanted to remain anonymous. "Once you get it down and can fill it out quickly and easily, it changes and they want different information than they were asking for before. It takes time away from what we actually get to do with people."[3]

No Analog to Profit

Because nonprofits are steeped in the daily struggle to help people, they usually prefer formative evaluations of their programs. In contrast, funders want to know, "What did we cause?" says Kuo, and therefore want summative evaluations. "But they don't understand that it's very expensive and difficult to set up

Selecting Best Practices

Too many nonprofits waste time and money reinventing the wheel. Instead of cobbling together a homemade intervention and trying to prove its effectiveness in the last few months of a funding cycle, many nonprofits would be better off utilizing programs that have already been proven effective. Several organizations review studies of program effectiveness and publish the results of their reviews online. Here are three online resources for best practices:

- The University of Colorado's Center for the Study and Prevention of Violence (CSPV) (www.colorado.edu/cspv/blueprints) has selected 11 "blueprint" programs for violence prevention that meet its standards for effectiveness.

- The Campbell Collaboration (www.campbellcollaboration.org) prepares systematic reviews of policies and practices in the areas of social welfare, education, and crime and justice. The Campbell Collaboration is modeled after the Cochrane Collaboration, an international nonprofit that systematically reviews healthcare interventions.

- The What Works Clearinghouse (www.whatworks.ed.gov), a project of the U.S. Department of Education, rates the effectiveness of educational interventions. The clearinghouse currently reviews programs in the areas of character development and middle-school mathematics.

evaluations that test causation." Kuo gives the example of a funder that wants to know whether its investment directly resulted in more kids graduating from high school. To answer this question, "you would have to follow students from their enrollment in the ninth grade until they graduated. And that would take four or five years," he notes, a time period that most funders wouldn't be willing to wait.

Indeed, the assumption that measuring nonprofit effectiveness is as quick and cheap as is measuring business performance frustrates many nonprofit leaders. "Board members dangerously assume that it might be as simple in this world as it is in business, but it isn't," says Phil Buchanan, executive director of the Center for Effective Philanthropy. "And it isn't even that simple in business," he adds.

"The next time corporate board members or donors get on an evaluation kick, ask them about the return on their investment in their R&D unit, or their advertising expenses," says Chip Heath, a professor at the Stanford Graduate School of Business. "They won't be able to tell you. And yet outcomes in the corporate world – which measures itself by profits – are much easier to measure than those that nonprofits are routinely asked to measure."

Patricia Patrizi, an evaluation consultant and chair of the Evaluation Roundtable, calls it the "*HBR* problem." "Everyone in the nonprofit world reads the *Harvard Business Review*, and now we're all trying to become corporate types," she says.

In fact, the social sector harbors no analog to profit – the quick and clean metric of the private sector. Success indicators for an arts organization in New York City are entirely different from those of a homeless shelter in Byhalia, Miss., a microlender in Bangalore, India, or an environmental advocacy group in the Amazon River basin. And because many innovative programs are only one step ahead of the issues they have been formed to address, it is not at all clear which indicators they should be tracking.

"What we know about innovative situations is that we don't even know what the appropriate targets are," notes Michael Quinn Patton, an independent evaluation consultant and former president of the AEA. "Yet the theory of change that dominates business approaches to foundations is that if we set targets and standards, that will make people meet them."

"Think about the civil rights movement," says Alan Durning, founder and executive director of Northwest Environmental Watch. "Many of the early civil rights workers in the South were trained in nonviolence by Quaker groups such as the Fellowship of Reconciliation. These groups provided some of the most powerful tools in the toolkit of a social movement that changed history. But at the time, did the Fellowship of Reconciliation know it would be a stimulus for the civil rights movement?" Quantifying social change is difficult work, Durning says, and often nonprofits cannot anticipate where their effects will be felt.

And had these history-changing nonprofits correctly foretold which outcomes to follow, could they have paid millions of dollars to track those outcomes over the decades it took for them to come to fruition? This question points out that social change is different from profit in another important way: It often runs at the speed of molasses. As a result, a program that looks like a failure at year two or three may prove a raging success at year 20 or 30.

For example, poor black children who enrolled in the High/Scope Perry Preschool intervention in Ypsilanti, Mich., in 1962 were not faring much better than their control group counterparts several years after they left the program, at ages 7 and 8. But some 40 years later, the High/Scope Perry gradu-

Agencies often do the secondhand Yugos of evaluation, rather than the Cadillacs that are required for good evidence, says Thomas Cook.

ates are more likely to have earned college degrees, have a job, own a home, own a car, have a savings account, be married, and have raised their own kids, reports David L. Kirp in *The New York Times Magazine*. They are also less likely to have been on welfare, to have been arrested, or to have been sentenced to prison.[4] Funders who might have decided to pull the plug on the High/Scope Perry program, based on early data, would have killed a program that has yielded over $12 on every $1 invested.[5]

Rocket Science

Boards and funders don't misuse evaluation terms and misapply business concepts because they are dumb or lazy, or even because they are ornery. Instead, their misuses and abuses of evaluation reflect the fact that good evaluation is an incredibly complex undertaking that most people do not understand. Even Ph.D.s and M.D.s who do evaluation for a living routinely make mistakes.

Take for example the many recent studies suggesting that a drink or two a day keeps the cardiothoracic surgeon away. On the surface, these studies seem solid. They have an experimental group – people who consume a couple of alcoholic beverages a day – and a control group of "abstainers" – people who don't touch the stuff. By comparing the treatment group to the control group, researchers can deduce whether drinking helps or hurts a person's health.

But a recent meta-analysis discovered a fatal flaw in 47 of the 54 studies: The researchers had not randomly assigned participants to their conditions. Instead, they grouped people who already drank two drinks per day into the experimental group and the people who already abstained from drink into the control group.[6] Lumped into this control group were people who had recently stopped drinking for medical reasons – especially alcoholism. In other words, the control group had more sick people than did the treatment group *to begin with*.

As a result, these studies did not test whether moderate drinkers are healthier than abstainers. Instead, they tested whether moderate drinkers are healthier than recovering alcoholics and other people too unhealthy to imbibe. And so it is no surprise that the treatment group fared better than the control group.

Seven studies avoided this mistake by excluding people who had recently gone off the sauce, so that their control group contained only the chronically abstemious. These studies showed that drinkers' health was no better than that of longtime abstainers – evidence against the drink-a-day hypothesis.

Why didn't the researchers randomly assign their participants and avoid this fatal flaw? For the same reason that many program evaluations don't use random assignment – it's extremely difficult and expensive. How would a cardiologist convince an oenophile to give up wine? How could she convince a teetotaler to take up drink? And how much would she have to pay participants to stick with their experimental regimen over the course of the study?

The complexities of random assignment – and, more generally, of the randomized control trial, which is what this kind of study is called – explode when the treatment moves out of the laboratory and into the real world. It takes a reliable IT infrastructure and highly skilled staff to track down enough participants, assign them to conditions, keep up with who is in what condition, make sure that the control participants aren't getting the treatment (you don't want the abstainers to shoot tequila on the sly), and make sure that the treatment participants are sticking to their medicines or programs (you don't want the drinkers to hop on the wagon or to slip into alcoholic dissipation). Add to these costs those of recruiting and compensating participants; entering and analyzing the data; and paying the overhead for all the people and facilities needed to do this, and randomized control trials quickly become prohibitively expensive. Very few funders are willing to tote this high a note.

Ethical Dilemmas

And then there's the ethical problem. Many social workers and nonprofit practitioners won't go along with the strictures of random assignment because they think it's morally wrong. If you believe that your program is the best thing for a drug addict, or a budding thespian, or an endangered estuary, why would you knowingly consign half of your recovery group, theater company, or wildlife species to a control group that offers what you think is second-best treatment? Many frontline service providers, when faced with this moral quandary, decide against using a control condition.

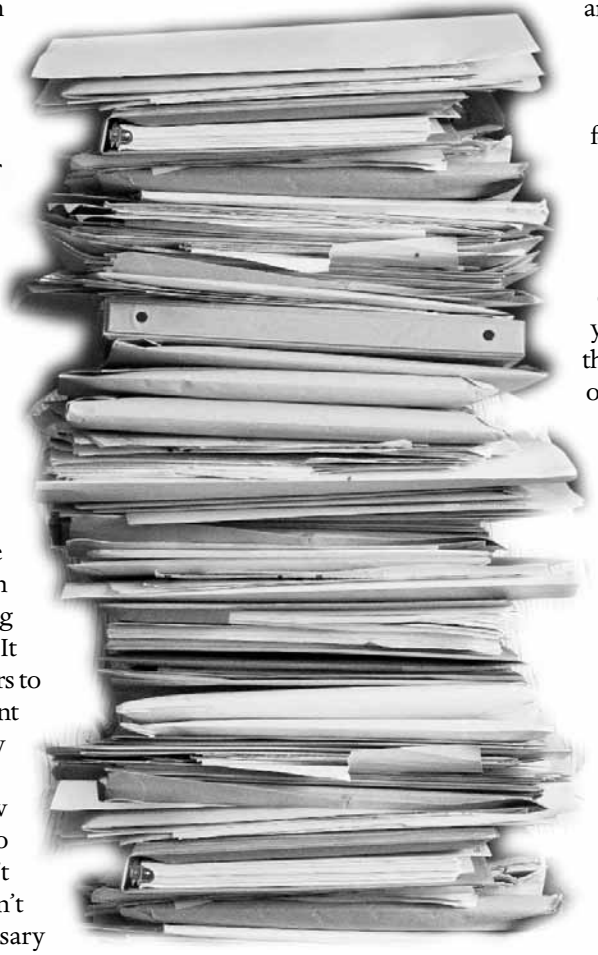
Not using a control group, however, can lead funders and policymakers astray, as Judith Gueron, former president of MDRC, demonstrated in her recent article in *Stanford Social Innovation Review*.[7] Looking only at the treatment group results – how many people in each of three programs left welfare and found jobs – you would naturally conclude that the program with the highest percentage was the most successful.

However, the treatment condition data don't take into account the fact that people often find jobs without the help of programs. You need a control condition to show how many people would have left welfare on their own. By subtracting the con-

Cutting out evaluation altogether is not an option. Instead, funders should ease nonprofits' burden by sharing it with them. The time has come for foundations to evaluate themselves.

trol condition outcomes from the treatment condition outcomes, you can then see what the program achieved above and beyond what people did on their own. Without a control condition, you can't separate how much of a program's success is due to its actions and how much is due to other factors (for example, a good economy).

Of course, the randomized control trial is not the only method for figuring out whether a program is working. But when it comes to inferring causation – that is, saying that A caused B or, in funders' parlance, that a program had an impact on a problem – the randomized control trial is the gold standard. And any departure from it – for example, not using random assignment, or not having a control group – opens the door for ambiguous results.



Down With Summative Evaluations

If there is one thing that funders, donors, and the general public can do to improve the situation, it's to stop insisting that nonprofits conduct summative evaluations. Ironically, fewer summative evaluations would result in better evidence. "One of the worst things a foundation can do is ask a grantee, 'How are you going to measure impact?'" says Patrizi. "It forces them to scramble for researchers to help them design some silly experiment that won't show much and that they can't afford."

Cook agrees: "Funders don't know how hard summative evaluation is to do. They don't know that you don't have the resources to do it. They don't know that if you hire in the necessary resources, it's coming out of program funds. And they don't appreciate that protecting program funds and the salary of the people in your program is the first duty of everyone who is running a service agency."

In these circumstances, agencies are forced to do what Cook calls a "secondhand Yugo" of an evaluation, rather than the Cadillac that is required for good evidence. "The problem with the secondhand Yugo is that it is not dependable. And in the future [unreliable data] has the potential to be an embarrassing mess."

Another reason to ditch summative evaluations is that they can be the enemies of innovation, says Patton. "Innovations by their nature are adapting solutions to conditions of uncertainty," he notes. "And so when you want people to be innovative, what you need is a process that allows you to try out things, alter things, change things, get feedback, and adapt." Summative evaluations require the opposite of innovation: deliver the same program over and over again, regardless of whether conditions change.

There are times and places for summative evaluations, but they are few and far between. Cook lists four criteria that the rare program should meet before subjecting it to a summative evaluation.

First, the program must have a decent theory of change. "Often when you ask people to lay out the theory of the program, six people will have six overlapping, but distinct theories of how it works," says Cook. Until everyone agrees on what a program is trying to do and how it is doing it, funders should not waste time and money measuring whether it worked.

Second, funders should also have evidence that program staff members are actually doing what they say they are doing – in other words, that they are implementing the program with fidelity. Third, funders should have reason to believe that the program is transferable elsewhere.

Finally, Cook says, "only evaluate proud programs" where people are

very pleased with and confident of what they are doing. Programs that do not meet these criteria, says Cook, should report on the theory of their program, their service delivery, and "maybe their outcomes, although I'm less keen on that."

TALK BACK: What are your reactions to this article? Post your comments at **www.ssireview.org.**

Sharing the Burden

As Aubry discovered, even when an organization bends over backwards to do a good evaluation, it does not necessarily help, because funding decisions often don't depend on data. Carolyn Roby has reached a similar conclusion from her perch as the vice president of the Wells Fargo Foundation Minnesota: "Big changes in funding strategy are not the result of unhappiness about the impact of previous grantmaking. It's just that someone gets a whim.

"Ten years ago in the Twin Cities, for example, employment was the big issue," says Roby. "Now, the sexy new thing is ready-for-kindergarten programs. It's not like our employment problems have been solved, or that our employment programs were bad. It's just that pre-K is hot, hot, hot. It drives me nuts."

The converse is also true – there are plenty of programs that are not proven effective, but that still bask in the warm glow of federal funding. DARE, which places police officers in classrooms to teach kids about the hazards of drug use, and abstinence-only interventions for teenage pregnancy have yet to show that they are better than – or even, in some cases, as good as – other programs.[8, 9] Yet DARE has been continuously funded for several decades, and abstinence-only programs show no signs of falling out of favor.

Since funders often do not use evaluation data in their decision making, it is all the more unfair that they foist it onto their grantees. But cutting out evaluation altogether is not an option. Instead, funders should ease nonprofits' burden by sharing it with them. The time has come for foundations to evaluate themselves, not just their grantees.

"Evaluating grant by grant will tell foundations about the individual grants, but creating real and sustained social change requires many actors," says Patrizi. Funders' responsibility should therefore be to build coalitions of grantees, and then to evaluate how effective the foundation as a whole is, rather than how effective each grantee's program is.

Funders should also coordinate with each other to consolidate their evaluation requirements, so that organizations like Rubicon don't have to create hundreds of different reports. One effort to do just that is already on the ground. The Center for What Works has compiled lists of the best outcome indicators for 14 program areas, such as adult education, performing arts, and prisoner reentry.[10] The center has also developed standardized logic models in each area. By using these standard measures and logic models, funders will not only lighten their grantees' evaluation load, but will also be able to compare programs more easily.

Another way that funders can improve evaluation is to partner with their grantees, turning evaluation into an opportunity for learning, rather than an occasion for judging. Evaluation scholars have developed several brands of these more learning-focused approaches to evaluation, such as Patton's utilization-focused evaluation,[11] or David Fetterman's empowerment evaluation.[12]

Both of these approaches use evaluation to create cultures of inquiry throughout the life cycle of a program. Funders' input is especially important at the beginning of a project, says Patrizi, because funders have broad experiences with a variety of programs. By spending more time and money thinking through a program's theory of change on the front end, rather than on collecting dubious data in the final months of funding, programs will have a better chance at success.[13] □

1 Carman, J.G. & Millesen, J.L. "Nonprofit Program Evaluation: Organizational Challenges and Resource Needs," *The Journal of Volunteer Administration* 23, no. 3 (2005).
2 Gammal, D.L.; Simard, C.; Hwang, H.; & Powell, W.W. "Managing Through Challenges: A Profile of San Francisco Bay Area Nonprofits," Stanford Project on the Evolution of Nonprofits (Stanford, CA: Stanford Graduate School of Business, 2005).
3 Quote comes from Gammal, Simard, Hwang, & Powell (2005). Because speakers participated in the research study with the understanding that their identity would be kept anonymous, this speaker's name is not given.
4 Kirp, D.L. "Life Way After Head Start," *The New York Times Magazine* (Nov. 21, 2004).
5 Belfield, C.R.; Nores, M.; Barnett, S.; and Schweinhart, L. "The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data From the Age-40 Followup," *Journal of Human Resources* 41, no. 1 (2006): 162-190.
6 Fillmore, K.M.; Kerr, W.C.; Stockwell, T.; Chikritzhs, T.; & Bostrom, A. "Moderate Alcohol Use and Reduced Mortality Risk: Systematic Error in Prospective Studies," *Addiction Research and Theory* 14, no. 2 (April 2006).
7 Gueron, J. "Throwing Good Money After Bad," *Stanford Social Innovation Review* 3, no. 3 (2005).
8 Ennett, S.T.; Tobler, N.S.; Ringwalt, C.L.; & Flewelling R.L. "How Effective Is Drug Abuse Resistance Education? A Meta-analysis of Project DARE Outcome Evaluations," *American Journal of Public Health* 84 (1994): 1,394-1,401.
9 Bennett, S.E. & Assefi, N.P. "School-based Teenage Pregnancy Prevention Programs: A Systematic Review of Randomized Controlled Trials," *Journal of Adolescent Health* 36 (2005): 72-81.
10 See http://www.whatworks.org.
11 See Patton, M.Q. *Utilization-Focused Evaluation,* 3rd ed. (Thousand Oaks, CA: Sage Publications, 1997); and Patton, M.Q. "Developmental Evaluation," *Evaluation Practice* 15 (1994): 311-320.
12 See Fetterman, D.M. *Foundations of Empowerment Evaluation* (Thousand Oaks, CA: Sage Publications, 2000).
13 Patrizi, P.P. "The Evaluation Conversation: Making Wise Decisions at Every Level of Foundation Practice," Paper in progress.