STANFORD
GRADUATE SCHOOL OF BUSINESS
STANFORD SOCIAL INNOVATION *review*

On the Frontlines

Throwing Good Money After Bad
A common error misleads foundations and policymakers

By Judith M. Gueron

Stanford Social Innovation Review
Fall 2005

STANFORD
GRADUATE SCHOOL OF BUSINESS
Stanford Social Innovation Review
518 Memorial Way, Stanford, CA 94305-5015
Ph: 650-725-5399. Fax: 650-723-0516
Email: info@ssireview.com, **www.ssireview.com**

Throwing Good Money After Bad

A common error misleads foundations and policymakers

by Judith M. Gueron

The program director of one of the nation's largest foundations and I were discussing a grant idea when our conversation turned to the topic of program evaluation. She revealed that she was under a lot of pressure from her foundation's trustees to show that the foundation's grants met their stated goals. While she strongly believed in the importance of demonstrating program effectiveness, she didn't have a clear idea of how to do it.

Returning to our grant idea, the director tried her hand at designing an evaluation, recommending that we report on outcomes, including how many people have a job and how many have completed high school by the end of the program. On one level, her suggestions seemed smart, reasonable, and doable.

But in fact she had fallen into a trap common among funders, service providers, and policymakers: She confused outcomes, an illusory index of effectiveness, with impacts – the real deal.

Outcome measures show the status of people at a point in time (e.g., before the program, upon leaving it, five years later), such as "55 percent of program graduates were drug-free six months after the program's completion" or "27 percent of program participants found stable housing within one year." Although these outcome measures seem to say something about effectiveness, they may not, because they fail to take into account one simple fact: Many people would have improved on their own, without help from the program. When programs rely on outcome measures to assess effectiveness, they have no way of differentiating how much of their clients' improvement is due to the program, and how much is due to clients' own actions.

As governor of Massachusetts, Michael Dukakis confused outcomes with impacts and mistakenly concluded that an educational program was effective.

So if a program serves many people – usually more advantaged – who manage to improve on their own, the program may be touted as a success. Conversely, if it serves many people – usually more disadvantaged – who are less likely to improve on their own, its outcome measures will be lower and it may be pronounced a failure.

The true measure of a program's effectiveness is therefore not its outcomes, but its impacts. Impacts are changes in outcomes that the program produces over and above what people would have accomplished on their own. To have an "impact" or to be "effective," a program has to get more people to work or to graduate or out of poverty than would have done so without the program.

Foundations are not alone in their confusion of outcomes with impacts. I recall hearing Vice President Quayle praising the high job placement rate of the Job Training Partnership Act that he had authored, and Governor Dukakis reporting that his Employment and Training Choices program

PHOTOGRAPH BY STEVE LISS/GETTY IMAGES

had moved large numbers of people in Massachusetts from welfare to work. Subsequent studies of both programs deflated these claims, suggesting that the majority of the job placements resulted from regular job-finding behavior, rather than from the special programs.[1]

Why Outcome Measures Mislead

A recent study of welfare reform programs in Riverside, Calif.; Atlanta; and Grand Rapids, Mich., shows how outcome data mislead.[2] Each of these programs sought to get single mothers off of welfare by requiring them to look for a job or to participate in some combination of short-term vocational training, adult education, or work experience. For the evaluation, people were randomly assigned to one of the following two groups: an *intervention group* that was enrolled in the site's welfare reform program, and a *control group* that was not. Since this study used a random assignment process – creating two virtually identical groups of people whose only difference was whether or not they were enrolled in a welfare reform program – its data can reliably show what the mothers could accomplish on their own (the control group's outcome), versus what more they could accomplish with the help of the program (the program's impacts).

JUDITH M. GUERON *is the former president of MDRC and is a visiting scholar at the Russell Sage Foundation. She is writing a book about the development of the randomized trial as a tool in the assessment of social programs and as an important factor in policymaking, focusing specifically on the case of welfare reform.*

Table 1 (p. 71) presents each program's outcomes and impacts for one measure: the percent of people working at some point in the second year after entering the study. Note that impacts are calculated by subtracting the control group's outcome from the intervention group's outcome.

The control groups' outcomes (first column of numbers) show how many people had found work without the welfare reform programs. The intervention groups' outcomes (second column of numbers) show how many people enrolled in their site's welfare reform program had found work.

The intervention groups' outcome data do not show, however, how many people were employed because of the program. These data are given in the fourth column of numbers – the programs' impacts.

These impact data hold the surprise, and the caution: Riverside, the site with the lowest intervention group outcome, had the highest impact, with an employment rate eight percentage points above that obtained by the control group. If a foundation had judged success based on outcomes alone, it would have funded the Grand Rapids and Atlanta programs, and not the slightly more successful Riverside one.

Why did the outcome data deceive? The answer lies in the fact that the three sites differ in more ways than just their welfare reform programs. They also have different labor market conditions, childcare opportunities, and kinds of welfare recipients. Riverside, for example, had higher unemployment during the time of the study, making it harder for people to find work, regardless of special program assistance.

Because outcomes jumble together the effects of programs with the effects of myriad other social, economic, and environmental factors, programs operating in strong labor markets or serving more advantaged people may have better outcomes, but actually accomplish less – that is, have less impact. Conversely, programs in weak labor markets or serving the less advantaged may make more of a difference – that is, have greater impact – but register lower outcomes.

This potential disconnect between outcomes and impacts, which is not unusual, raises a red flag for foundations. It doesn't mean that achieving high outcomes is irrelevant, but it does mean that basing funding decisions on outcome measures alone can be dangerous. This practice may not only cause foundation staff to reward ineffective programs or to neglect effective ones; it may also prompt programs to maximize their outcomes by changing whom they serve, rather than by investing in higher-quality services.

Is Measuring Impacts Worth the Trouble?

Evaluators agree that the most reliable way to measure a social program's impact is the randomized controlled trial, which is the same method that medical researchers use to assess a new product or procedure's effectiveness. In this method, evaluators use a lottery (called "random assignment") to create two groups: one that is enrolled in the program and one that is not.

Because people are randomly assigned to groups – instead of assigned according to some other criterion or allowed to choose – the groups are assumed to be the same on all other social, economic, and environmental dimensions *besides* the program. Thus

PHOTOGRAPH BY SHANNON FAGAN

Table 1
Programs With Low Outcomes May Have High Impacts

Program	OUTCOMES			IMPACTS	
	Control Group	Intervention Group			
	% Working	% Working	Rank	% Working	Rank
Grand Rapids	61	67	1	6	2
Atlanta	53	57	2	5	3
Riverside	38	46	3	8	1

Note: Due to rounding, impacts may not exactly equal the difference between the intervention and control groups' outcomes.

SOURCE: Unpublished MDRC data

evaluators can conclude that any differences they see between the two groups are due to the program's influence, and not to other factors.

While other evaluation methods can sometimes be convincing and appropriate, they more commonly lead to uncertainty or the wrong conclusions. The usual result is: We can't tell.[3] At the same time, randomized studies, while often feasible, can be demanding and expensive. Are they worth it?

In my experience, the more unimpeachable the evidence, the greater the likelihood that an evaluation will be seen as believable, and not just as noise from yet another pressure group. For example, former Senator Daniel Patrick Moynihan stated that the Ford Foundation-sponsored studies of state welfare reform initiatives shaped his own welfare reform legislation. His senior staffer attributed the studies' influence in large part to their "rigorous methodology – experimental design with random assignment."[4] Congress later asked the U.S. Department of Health and Human Services to use randomized studies to assess both the 1988 and 1996 welfare reform bills.

Arguments for education reforms have likewise heavily drawn on rare randomized studies. Proponents of prekindergarten programs have cited the Perry Preschool and Abecedarian studies, and states have used the Tennessee STAR experiment to argue for reduced class size. Just as importantly, rigorous studies of unsuccessful programs for disadvantaged high school dropouts have moved that field away from dead ends and toward more promising innovations.[5]

When and How to Measure Impacts

Despite the importance of measuring impacts, no foundation can do this for all of its programs. Here are some guidelines for when and how to approach effectiveness evaluations:

1) Discuss the meaning of effectiveness with trustees, staff, and grantees. This will encourage the collection of more meaningful data and discourage the "cherry-picking" of programs and clients that would show high outcomes.

2) Fund programs that adapt and replicate strategies that high-quality studies have already proven effective.

3) If no rigorous studies exist in a central program area, advocate for a study that will address the big questions. This study should not only measure the programs' impacts, but should also explore the management strategies and service practices that lead to superior or poor performance. You probably will not be able to launch this study alone, and will thus need like-minded collaborators. When you seek to sell the evaluation, focus not only on its direct cost, but also on its long-term potential to influence policy.

4) Resist pressure to assess the effectiveness of all grants or programs. Do not be embarrassed to say that an evaluation would be too expensive if done adequately or a waste of money if done poorly.

Whatever the program area, identifying and promoting policies and practices that work are fundamental to making things better. So is shedding those that do not. If we can accurately assess programs' effectiveness, we have the potential to shift resources toward programs with the greatest impact, to defend successful activities from assault, to preserve or increase funding in the areas we care about, and ultimately, to help more people in more useful ways. □

1 See, for example, L.L. Orr, H.S. Bloom, S.H. Bell, F. Doolittle, W. Lin, and G. Cave, *Does Job Training for the Disadvantaged Work? Evidence From the National JTPA Study* (Washington, D.C.: Urban Institute Press, 1996).

2 For more on this evaluation, see G. Hamilton, *Moving People From Welfare to Work: Lessons From the National Evaluation of Welfare-to-Work Strategies* (Washington, D.C.: U.S. Department of Health and Human Services and U.S. Department of Education, 2002).

3 For a summary of this literature, see Chapter 5 in H.S. Bloom, ed., *Learning More From Social Experiments: Evolving Analytic Approaches* (New York: Russell Sage Foundation, 2005) and L.L. Orr, *Social Experiments: Evaluating Public Programs With Experimental Methods* (Thousand Oaks, CA: Sage Publications, 1999).

4 E.B. Baum, "When the Witch Doctors Disagree: The Family Support Act and Social Science Research," and P.L. Szanton, "The Remarkable 'Quango': Knowledge, Politics, and Welfare Reform," both in *Journal of Policy Analysis and Management* 10 (1991): 590-615.

5 See Orr et al, (1996) and J. Bos, G. Cave, F. Doolittle, and C. Toussaint, *Jobstart: Final Report on a Program for School Dropouts* (New York: MDRC, 1993).