The Role of Nonprofits in Designing and Implementing Evidence-Based Programs

James X. Sullivan

3108 Nanovic Hall, Wilson Sheehan Lab for Economic Opportunities, Department of Economics, University of Notre Dame, Notre Dame, IN 46556. Phone: (574) 631-7587. Email: jsulliv4@nd.edu

James X. Sullivan is the Rev. Thomas J. McDonagh, C.S.C., Associate Professor of Economics at the University of Notre Dame, and the director of the Wilson Sheehan Lab for Economic Opportunities. His current research focuses on measuring the effect of domestic anti-poverty programs.

NOTE: This work was supported by the Wilson Sheehan Lab for Economic Opportunities. I am grateful to Wendy Barreno, Rachel Fulcher-Dawson, and Katie Kuka for their extremely helpful advice. I also benefitted from thoughtful comments from Ron Haskins, Tom Sheridan, and participants at the Brookings Authors' Conference.

Abstract

Human service nonprofits are a major provider of social services in this country, spending billions of dollars each year implementing programs to improve outcomes for their clients. Unfortunately, these programs are typically not rigorously evaluated to determine whether they are having their intended effect. Many obstacles make it challenging to rigorously evaluate services provided by these nonprofits, including evaluation costs, limited access to data, and small sample sizes, but these obstacles are surmountable. Policymakers could accelerate the pace and quality of evidence-building by providing more resources for impact evaluations, streamlining and standardizing access to key administrative data, and expanding support for the

replication of effective programs. Better evidence of what works for human service nonprofits will ultimately mean more effective programs at the national level.

Keywords: non-profit; impact evaluation; evidence-based policymaking; social programs

In recent years policymakers, researchers, and advocacy groups have increasingly emphasized evidence-based policy making; that is, the idea that policy decisions should be shaped by rigorous evidence.[1] Congressional commissions have addressed the topic, and books have highlighted its emergence.[2] More than ever, government agencies are relying on evidence for decisions about how to allocate resources through merit-based grants. Proponents of evidence-based policy making argue that if we spend scarce resources on effective programs, public and private funds would have a much larger impact on the lives of people living in poverty.[3]

Federal, state, and local governments spend billions of dollars each year on social programs that affect the lives of millions of individuals and families. Many of these programs are implemented at the local level, often by human service nonprofits.[4] If these programs are not evidence-based, then the policies that support them won't be, either. As the co-chair of the Commission on Evidence-Based Policymaking, Ron Haskins, put it: "In my way of thinking about evidence-based policy, the single most important player is the group of people who establish and run programs that actually deliver services at the local level. All politics might not be local, but all program implementation is." Unfortunately, we do not know what works at the local level, because most promising social programs are not evaluated rigorously. Rigorous evidence does not play a prominent role in major program, funding, and scale-up decisions. Many human service nonprofits lack access to the resources and data to measure impact. Overcoming these barriers to evidence building at the local level would go a long way toward promoting evidence-based policymaking at the national level.

Human Service Nonprofits and their Role in Evidence-Based Programming

Human service nonprofits offer a variety of services, including job training, crime prevention, nutrition assistance, affordable housing, youth development, foster care, disaster relief, and many other essential programs. According to the National Center for Charitable Statistics, revenue for these providers exceeds $200 billion annually.[5] Human service nonprofits receive funding from many sources, including private charitable contributions, government grants and contracts, and fees for goods and services. In 2012, governments contracted with these providers for about $81 billion (Urban Institute 2015).

Human service nonprofits offer an ideal breeding ground for innovative, evidence-based social programs that can be replicated nationally. That's because to know what to replicate, we first need to know what works on a small scale. Equally important, ideas for new programs typically come from the local level, often inspired by a community issue. For example, large numbers of disconnected youth (those who are neither working nor in school) might spark a youth employment program, or rampant homelessness might lead to a housing program. The ideal time to measure impact is when a program is in its early stage. If a scientific evaluation shows that it is effective, that evidence will help raise additional resources so that it can be replicated and scaled up.[6]

Many large, national programs were designed and implemented at the local level and scaled up precisely because they were shown to be effective. For example, the Nurse-Family Partnership (NFP), a home visitation program for new, low-income mothers, started as a small intervention in Elmira, NY. After several randomized controlled trial (RCT) evaluations showed that the program improved outcomes for both mothers and children, the NFP has been scaled up and now serves more than 32,000 families in 42 states (Haskins and Margolis 2014, chapter 2).[7]

Its documented success is a primary reason the federal government has invested millions of dollars in the NFP and other home visitation programs.

Unfortunately, the NFP is more the exception than the rule. Unlike medicine and, increasingly, business, social services do not have a strong culture of rigorous impact testing. Most providers design and implement programs with little to no hard evidence that they work, and local programs are often replicated and scaled up without knowing their true impact. Only 8 percent of nonprofits have an evaluation staff (Beer 2016), and only 2 percent have conducted an RCT (Veris, 2013).

Moreover, once these programs are implemented on a large scale, it becomes hard to scale them back even if new evidence shows that they are not having the intended impact. For example, Drug Abuse Resistance Education (DARE) was designed to respond to youth drug abuse problems that emerged and grew in the 1970s and 1980s (Cima 2016). DARE was a 17-week curriculum administered by police officers that focused on decision-making skills and resisting drug use. Launched in Los Angeles, it quickly spread across the nation. At its high point, 75 percent of US schools used the curriculum, including schools in all 50 states (schools in 52 other countries used it as well) (Nordrum 2014). Yet a series of evaluations that followed up with students at one, five and 10 years after the program demonstrated that it had no effect on drug use (Ennett et al. 1994; West and O'Neal 2004). Nearly two decades after the program began, the federal government defunded DARE and denounced it for its null and occasionally even negative effects (GAO 2003; Surgeon General 2001).

Barriers to Evidence-Based Programming

If human service nonprofits are the ideal place to foster evidence-based policy, why do we see so little rigorous evidence in this sector? Two things stand in the way. First, practical barriers—such as evaluation costs, small sample sizes, and limited access to data—make it hard or even impossible to produce rigorous evidence of program impact. Second, conventional impact evaluations often fail to suit the needs of the provider. Although these obstacles are substantial, they are far from insurmountable.

Service providers face many practical barriers when they aim to measure program impact, perhaps the most important of which is limited resources. Providers often struggle to secure the funds they need to implement programs that meet the community's needs, so spending money on evaluations can be a hard sell. It can be particularly difficult to raise the money to rigorously evaluate a new program that has yet to establish evidence of promise. Providers may not see how they can measure outcomes at a reasonable cost and on a reasonable timeline, and therefore may not request evaluation funds from government grants, philanthropy, and other funders.

Moreover, government and private funders often do not support evaluations, and those that do rarely require rigorous research designs that measure a program's causal impact on key outcomes. Most foundations include funding for evaluations with fewer than 10 percent of their grants (Beer, 2016). Government and private grants that require evaluations typically only ask grantees to track certain outputs (i.e. how many people were served) or to measure some basic outcomes for those who receive services. Even when funders include strict requirements for evaluations, the quality of the evidence may not be strong. For example, the Social Innovation Fund (SIF) has spent hundreds of millions of dollars supporting local nonprofit programs that are

backed by evidence. However, of the 77 evaluations approved by 2014, only about a third were RCTs. About half were quasi-experimental studies (evaluations that are similar to RCTs but that do not rely on random assignment to define the treatment and control groups), but many of these did not have rigorous research designs (Haskins and Margolis 2014, p. 165).

In many instances, a rigorous research design is simply not feasible because the program serves only a small number of people. Small sample sizes limit the power of statistical tests, making it difficult to measure a program's effects precisely, even when the true effect is large. In other instances, even though a program serves a large number of clients an impact evaluation still may not be feasible because the demand for services is not sufficient to generate an appropriate comparison group. If a job training program serves all eligible applicants, for example, no potential clients are left for the control group. Programs without excess demand for their services among an eligible population can still measure program impact through quasi-experimental approaches that compare eligible clients to ineligible clients, but some of these methods require even larger sample sizes, and providers are often not aware of these options.[8]

Even when funding is available to evaluate program outcomes, social service providers may not collect or have access to the data they need to do so. Many providers collect information about the services they provide their clients, but gathering information on clients' outcomes is less common, because the appropriate time to measure impact is typically after the client has completed the program. And few service providers have access to data on outcomes for a comparison group of people who do not receive their services, which they would need to isolate the program's impact.

Collecting outcome data for an evaluation can be expensive. In-person follow-up surveys, for example, can cost upward of $500 per person. The good news is that, in many instances,

surveys are not necessary because administrative records contain information on key outcomes such as employment, earnings, college persistence and completion, contact with the criminal justice system, and hospital admissions, among others. But social service providers and their research partners typically can't access these data for impact evaluations.

Even when none of these common practical barriers is an issue, rigorous evaluations do not always suit the needs of the provider. If the key outcomes in an impact evaluation won't occur until far in the future, providers can't use the information in the short term to raise funds or improve programing. For example, a program that promotes success in college may have to wait four or more years to measure its impact on a key outcome such as college graduation. In addition, putting programs under a microscope can be daunting. Evaluations can and often do show that programs are not having their intended effect. Moreover, when programs with solid evidence of impact are replicated, they often do not produce the same promising results (Baron and Sawhill 2010).

Another limitation is that RCTs are not usually designed to determine which features of an intervention are most and least effective. In other words, RCTs are better at telling us what works than why it works. The well-known RCT evaluations of the Perry Preschool project produced rigorous evidence that the intervention had a positive impact on a number of long-term outcomes (Heckman et al. 2013; Schweinhart et al. 2005). But the Perry Preschool Study does not tell us which components of the intervention—the curriculum, the home visits, the teachers, etc.—were critical in producing its effects. RCTs can measure the impact of specific program components if they randomly assign clients into multiple variations of a program, but such studies require much larger samples than an RCT evaluation that measures overall program impact.

Promoting More Evidence-Based Programming

Overcoming the obstacles reviewed above is essential to promote evidence building among human service nonprofits, but some of the practical barriers to impact evaluations will be hard to break down. For example, it will always be hard to measure the impact of small programs and those with little excess demand. Such programs should be encouraged to build evidence of promise through non-experimental means—for example, by comparing outcomes for their clients to those of a comparison group constructed to match the clients' demographic characteristics (Blundell and Costa Dias 2000). Evidence of promise can help providers raise the funds they need to scale up small programs so that they're big enough for a full experimental evaluation. In some cases, programs that lack excess demand can increase interest by promoting their services more broadly, or by expanding eligibility. Or they may measure the impact through quasi-experimental methods.

Because evaluation is costly, governments and private foundations need to better support rigorous evaluation of promising programs. A recent example of how government agencies can encourage evidence-based programming is the Department of Education's Investing in Innovation (i3) initiative, which has used a tiered-evidence model to distribute more than $1 billion in grants to improve student achievement.[9] In the tiered-evidence approach, funds are allocated by merit-based competitions, as opposed to formula grants where geography or other factors are more important than rigorous evidence. The lowest tier ("development") i3 grants support promising initiatives that lack rigorous evidence. These grants create a pipeline for innovative programs that, if proven effective, can be scaled up for broader impact. The top tier ("scale-up") i3 grants go to initiatives with one or more well-designed and implemented RCTs or quasi-experimental studies.

Sometimes legislation creates funding to test new programs. For example, Section 4022 of the 2014 Farm Bill authorized $200 million to support 10 pilot projects designed and implemented by state agencies to reduce the need for public services and encourage employment among Supplemental Nutrition Assistance Program (SNAP) participants. Each pilot project was required to have an independent evaluation that compared outcomes for households participating in the pilot to a control group of households not participating. The legislation also required states to make available administrative data that could be used to track outcomes. More pilot initiatives like this could go a long way toward promoting better evidence-based programming.

Service providers and their research partners often lack access to administrative data sources that could give them information about outcomes. If they could link micro-level data on program participants (and a comparison group) to administrative data, they would have much more evidence to assess program impact. Often, the best comparison group is drawn from program applicants who are eligible for a program but cannot be served because of insufficient capacity. If they could routinely collect administrative data on program applicants who are eligible but not served, providers would be better able to measure program impact in a fairly rigorous way in both the short and long run. Ideally, service providers would have clear and consistent protocols for linking their program data to administrative data on outcomes. Such protocols would include a standardized list of personal identifying information (name, date of birth, etc.) that would allow them to link to these data sources. Providers would also benefit from a standardized process for protecting privacy when linking and sharing data, including clear and consistent protocols for informed consent.

Administrative datasets that would be particularly useful to social service providers include earnings records, use of public benefits, hospitalization and other health outcomes, health

care use, arrest records and other criminal justice information, credit reports, and education records. Giving providers and their research partners access to these data would make impact evaluations easier to conduct, encourage more researchers to focus on policy-relevant studies, and give policymakers better evidence of program impact and effectiveness—leading to more effective social service programs and policies and, in turn, improved outcomes for program participants.

Promoting evidence-based programming requires more than just producing strong evidence of effective programs. Providers need to act on the evidence so that the best programs are scaled up and replicated. Many social service providers do not have ready access to information about the most effective programs. Even when such information is available, it can be hard for providers to sift through many studies or to separate strong from weak evidence. And information about how to successfully replicate evidence-based programs can be hard to come by.

To design and implement evidence-based programs, social service providers need a way to track down and navigate the evidence on what works best, for whom, and under what circumstances. A national repository of well-designed, well-implemented impact evaluations would help to promote a broader culture of continuous evaluation and improvement. Several clearinghouses with important information about effective programs already exist, and they can serve as a foundation.[10] One important challenge for a national repository will be to help stakeholders by offering clear standards for what constitutes reliable evidence. Ideally, an independent entity would assess evaluations and identify the reliable ones.

But reliable evidence alone does not ensure that effective programs are implemented broadly, because providers typically do not have access to information on how to successfully

replicate these proven programs. Thus, providers also need guidelines and funding to replicate evidence-based programs in new settings and/or with new target audiences, including resources for continuous evaluation and improvement. Such support would help to ensure that the most effective programs are implemented broadly and with fidelity. Without fidelity, it will be harder to reproduce the results and impact of the original program.

If rigorous evaluations are to become more common, evaluators will need to find ways to overcome the problem that evaluations do not always suit a provider's needs. One way is to produce more timely evidence that providers can readily use to improve their programs rather than simply to evaluate overall impact. More accessible data will help make useful, near-real-time evidence possible.

To encourage more providers to embrace evidence-based programming, an impact evaluation should not be characterized as a high-stakes assessment but rather as an important diagnostic tool to continuously improve programming (Haskins and Margolis 2014, p. 234). That can be a problem when impact evaluations are conducted on programs that are already at scale, because by that time the intervention model is often well established. However, impact evaluations can show how to improve the effectiveness of new, smaller-scale programs (see chapter ??? by Knox, Hill, and Berlin). In a broad culture of evidence-based programming—where interventions are continuously evaluated, the evidence from those evaluations leads to improved programming, and resources are allocated to programs that work best—ineffective programs will be less likely to reach the point where they are evaluated at scale. That is, measuring the impact of new programs will make it more likely that only the best programs are tested on a large scale.

Conclusion

Understanding what works at the local level will lead to more evidence-based programs at the national level. Though substantial obstacles stand in the way of generating rigorous evidence, these obstacles are surmountable. Policymakers could accelerate the pace and quality of evidence-building by providing more resources for impact evaluations, streamlining and standardizing access to key administrative data, and expanding support to replicate effective programs. These steps would make evidence-based programs significantly more common at all levels of government, so that scarce resources are allocated to the most effective programs.

References

Angrist, Joshua D., and Victor Lavy. 1999. Using Maimonides' rule to estimate the effect of class size on student achievement. *Quarterly Journal of Economics* 114 (May): 535-575.

Baron, Jon, and Isabel Sawhill. 1 May 2010. Federal programs for youth: More of the same won't work. *Brookings*. Available from www.brookings.edu.

Beer, Tanya. 2016. Evaluation demand and capacity in the social sector. Presentation before the Commission for Evidence-Based Policymaking, 4 November 2016. Washington, DC. Available from https://www.cep.gov/content/dam/cep/events/2016-11-04/114Beer.pdf.

Blundell, Richard, and Monica Costa Dias. 2000. Evaluation methods for non-experimental data. *Fiscal Studies* 21 (4): 427–468.

Cima, Rosie. 19 December 2016. DARE: The anti-drug program that never actually worked. *Priceonomics*. Available from https://priceonomics.com/dare-the-anti-drug-program-that-never-actually/.

Ennett, Susan T., Nancy S. Tobler, Christopher L. Rigwalt, and Robert L. Flewelling. 1994. How effective is drug abuse resistance education? A meta analysis of Project DARE outcome evaluations. *American Journal of Public Health* 84 (9): 1394–1401.

GAO. 15 January 2003. *Youth illicit drug use prevention: DARE long-term evaluations and federal efforts to identify effective programs.* Washington, DC: United States General Accounting Office. Available from http://www.gao.gov/assets/100/91676.pdf.

Haskins, Ron, and Greg Margolis. 2014. *Show me the evidence: Obama's fight for rigor and results in social policy*. Washington, DC: Brookings Institution Press.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics* 94 (1): 114-128.

Nussle, Jim and Peter Orszag. 2014. Let's play moneyball. In *Moneyball for government*, eds. Jim Nussle and Peter Orszag, 2-11. United States: Disruption Books.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores. 2005. *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Press.

Surgeon General. 2001. Chapter 5: Prevention and intervention. In *Youth violence: A report of the Surgeon General*. Washington, DC: Dept. of Health and Human Services, U.S. Public Health Service.

Veris. 2013. *Executive Summary on the State of Scaling Among Nonprofits* (New York: Veris Consulting and the Social Impact Exchange).

West, Steven L., and Keri K. O'Neal. 2004. Project D.A.R.E. outcome effectiveness revisited. *American Journal of Public Health* 94 (6): 1027–1029.

Notes

[1] While evidence can take on many forms, the call for evidence-based policymaking emphasizes the importance of rigorous and objective measurement of program impact. Randomized controlled trial evaluations are the gold standard for rigorous evidence, but other quasi-experimental approaches are often viewed as providing solid evidence of program impact (Blundell and Costa Dias, 2000).

[2] See https://www.cep.gov/ and Haskins and Margolis (2014) or Nussle and Orszag (2014).

[3] For example, see Speaker Ryan's Press Office, "Speaker Ryan Names Appointees to Evidence-Based Policymaking Commission," June 2016. http://www.speaker.gov/press-release/speaker-ryan-names-appointees-evidence-based-policymaking-commission, or Senator Patty Murray's Press Office, http://www.murray.senate.gov/public/index.cfm/newsreleases?ID=B402B72B-547C-47EB-83B7-E64BFDD8A2EF.

[4] This article focuses on human service nonprofits (or service providers), but many of the points also apply to other nonprofits such as those working in education or health, and to state and local governments that implement social programs.

[5] Total revenue in 2013 for the human services sector was estimated to be $214 billion; see Urban Institute, National Center for Charitable Statistics, Core Files (Public Charities, 2013). http://nccsweb.urban.org/PubApps/showDD.php#Core%20Data.

[6] Even when rigorous evaluation shows that a local program is effective, this does not mean that the program will work on a larger scale or in a different community with different clients facing different needs.

[7] http://toptierevidence.org/programs-reviewed/interventions-for-children-age-0-6/nurse-family-partnership.

[8] One common approach, for example, is the regression discontinuity research design that compares outcomes for groups on either side of an eligibility threshold. For example, see Angrist and Lavy (1999).

[9] In 2015, the Every Student Succeeds Act replaced i3 with a similar program called Education Innovation and Research.

[10] For example, the U.S. Department of Education's research arm—the Institute of Education Sciences (IES)—runs the What Works Clearinghouse.