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

James X. Sullivan
University of Notre Dame and
Wilson Sheehan Lab for Economic Opportunities

June 4, 2017

I. Introduction

In recent years policymakers, researchers, and advocacy groups have placed increasing emphasis on evidence-based policymaking—the idea that policy decisions should be shaped by scientific evidence. Congressional commissions have been launched to address the topic and books have been written highlighting its emergence.[1] More than ever before, government agencies rely on evidence for decisions about how to allocate resources through merit based grants. Proponents of evidence-based policy making argue that by allocating scarce resources towards proven programs, the government can do more to improve the lives of those in need.[2]

Federal, state, and local governments spend billions of dollars each year supporting social programs that impact the lives of millions of individuals and families across the country. Implementation of many of these programs occurs at the local level, often by human service nonprofits. Therefore, these organizations play an important role in shaping social policy, and in order to design social policies that are informed by evidence, we must understand the impact of these local social service programs.[3] As the co-Chairperson for 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. 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 typically not evaluated rigorously.[4] Hard evidence does not play a prominent role in which programs are scaled up and which ones receive government funding. This evidence is lacking in large part because many human service nonprofits lack access to the resources and data necessary to measure impact. Addressing these barriers to evidence at the local level would go a long ways towards promoting evidence-based policymaking at the national level.

---

[1] See https://www.cep.gov/ and Haskins and Margolis (2014) or Nussle and Orszag (2014).
[2] 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.
[3] This article focuses on human service nonprofits (or service providers), but many of the points made here also apply to other nonprofits such as those working in education or health, and to state and local governments that implement social programs.
[4] 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).

II. 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 to communities across the country. According to the National Center for Charitable Statistics (NCCS), these programs amount to more than $200 billion annually.[5] Human service nonprofits receive revenue 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 the ideal breeding ground for innovative, evidence-based social programs that can be replicated nationally. To know what to scale up, we need to know what works on a small scale. Ideas for new programs typically come from the local level, often inspired by a community issue. For example, high rates of teen idleness might spark a youth employment program, or rampant veteran homelessness might lead to a supportive housing program. If the program is implemented well and shows some evidence of promise, it might receive additional funding to expand. This early stage in a program's development is an ideal time to measure its impact. If a scientific evaluation shows that the program is effective, this solid evidence will help raise additional resources so that the program can be replicated and scaled up.[6] If subsequent evaluations reinforce the initial positive findings and there is sufficient need for the intervention, then opportunities to expand the program will continue as other providers seek to implement and policymakers seek to support evidence-based programs.

Many large, national programs were initially 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. Backed by several randomized controlled trial (RCT) evaluations showing 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] The impact of this evidence goes beyond the scale-up of the NFP. The documented success of this program is a primary reason the federal government has invested millions of dollars in other home visitation programs.

Unfortunately, the NFP is more the exception than the norm. Unlike medicine, and increasingly the business world, there is not a strong culture of rigorous testing of impact in the social services. Most providers design and implement programs with little to no hard evidence of

---

[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] It should be noted, that 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. In other words, the results from local programs are not necessarily generalizable. When a proven program expands or serves a different community with different clients facing different needs, the impact of the program may change. This is an important reason why the results from many rigorous studies are hard to replicate. This is also why it is important to continuously evaluate a program as it is expanded and replicated to verify its impact in different contexts.

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

impact. Only 8% of nonprofits have an evaluation staff (Beer, 2016). Through government initiatives, local programs are often replicated and scaled up with little or no evidence on program effectiveness. Moreover, once these programs are implemented on a large-scale, it becomes difficult to scale them back, even if new evidence shows that they may not be having the intended impact.

For example, the Drug Abuse Resistance Education (D.A.R.E.) program was designed to respond to youth drug abuse problems emerging in the 1970s and 1980s (Cima, 2016). DARE was a 17-week curriculum to be administered by police officers that focused on decision-making skills and resistance. Launched initially in Los Angeles, it quickly spread nationally – at its high point it was in 75% of all U.S. schools in all 50 states (as well as 52 other countries) (Nordrum, 2014). The Drug-Free Schools and Communities Act of 1986 set aside funding for states that did drug prevention education in schools and specifically named DARE as one such program (Cima, 2016). Subsequently, a series of evaluations demonstrated no effect on drug use – following students 1, 5 and 10 years (Ennett, et al. 1994; West and O'Neal, 2004). Nearly two decades after the initial locally grown program began, the federal government defunded DARE and decried it for its null and negative effects (GAO, 2003; Surgeon General, 2001).

The Even Start Literacy Program offers another example. In 1986, the Kentucky legislature funded the Parent and Child Education (PACE) program, an initiative designed to improve both child and parent literacy. Without being rigorously evaluated, the program quickly expanded to 30 counties.[8] Modeled off PACE and similar programs, in 1989 the federal government launched the Even Start Literacy Program, allocating $14.5 million to support 76 demonstration grants (U. S. Department of Education, 2003). Three national evaluations showed that the program had little impact (U. S. Department of Education, 2003). Even after the release of these findings, more than $1 billion was allocated to the program, and it was more than 10 years before resources were redirected to alternative interventions (Bridgeland and Orszag, 2013).

Government agencies rely heavily on human service nonprofits to deliver a variety of essential social services, and this dependence has grown significantly over time.[9] Because public grants and contracts are a significant funding source for providers, government decisions about how to allocate resources for social services greatly affects the nature of the programs that these local organizations offer. Government agencies often issue a grant or notice of funding availability (NOFA) inviting non-profit organizations to apply for funding to implement a specific program. For example, in 2016 the U.S. Department of Housing and Urban Development issued a NOFA for $1.9 billion to support the Continuum of Care Program that provides homelessness services at the local level.[10] Organizations responding to the notice could construct their own programs, but the NOFA included a long list of program requirements. Funding decisions are not made based on the extent to which the program is backed by hard evidence. And although awardees are asked to track outcomes, they are not required to rigorously evaluate their work.

---

[8] https://wvde.state.wv.us/abe/wvfli/four_components.html.

[9] In 1997, about 52 percent of all government spending on social services went to nonprofits (including human service organizations as well as other nonprofits such as health and education, Urban Institute, 2010; Solomon 2003).

[10] https://www.hudexchange.info/resources/documents/FY-2016-CoC-Program-NOFA.pdf.

III. The Lack of Evidence-Based Programming

If human service nonprofits are the ideal place to foster evidence-based policy, then why do we see so little rigorous evidence in this sector? One explanation is that generating hard evidence often does not align with the primary goals of service providers. However, even when a rigorous impact evaluation suits the needs and interests of a provider, there are often practical barriers that make it quite challenging, or even impossible, to produce scientific evidence of program impact.

Scientific evaluations are an extremely powerful tool that can provide organizations with critical information on program impact and offer the kind of evidence necessary to scale up a program to have broad, national impact. The primary goal of most providers, however, is to offer services that address an important need for a disadvantaged population in a local community, not to produce evidence to inform policymakers or other providers at a national level. Furthermore, impact evaluations often do not provide specific information on how to improve a program. The time-to-evidence for many impact evaluations is slow, particularly if the key outcomes of interest are far off in the future. For example, a program designed to promote college graduation rates among new enrollees, will not have information on key outcomes for at least four years. This limits the providers' ability to leverage the information from an evaluation to raise resources or improve programing.

RCTs are typically designed to measure overall program impact—the effect of a package of services offered to the treatment group but not the control group. These impact evaluations provide the best evidence possible on the extent to which the program is affecting key outcomes. However, RCTs are not usually designed to determine which specific features of an intervention are the most effective, or which features do not move the needle enough to justify the investment. In other words, RCTs are better at telling us what works than why it works. The well-known RCT evaluations of the Perry Pre-school project provided rigorous evidence that the intervention had a positive impact on a number of long-term outcomes (Heckman et al., 2013; Schweinhart et al., 2005). However, the Perry Study does not tell us which components of the intervention—the curriculum, the home visits, etc.—were the most critical for moving the needle. Evaluators can design RCTs to measure the impact of specific program components by randomly assigning clients into multiple treatment arms that vary by which components of the intervention are offered, but such studies are more complicated to implement and require much larger samples than an RCT evaluation that measures overall program impact.

Some providers may be reluctant to put their programs under a microscope in fear of what we might find. These concerns may be well-founded; many large-scale impact evaluations show that programs are not having their intended effect, and replications of programs with solid evidence of impact often do not produce the same promising results (Barron and Sawhill, 2010). On the other hand, most providers are deeply invested in their programs, and consequently are often confident that their programs are making a difference.

Practical barriers—including cost, program capacity, demand for services, and limited access to data or information about alternative research designs—tend to be the most important obstacles service providers face when aiming to measure program impact. Many social service providers

have limited access to the resources necessary to measure impact. Designing and implementing an impact evaluation for a new intervention can be particularly challenging because it is difficult to raise resources to evaluate an unproven program (Haskins 2014 p. 135). Providers may not see opportunities to measure outcomes at reasonable costs and on reasonable timelines and, therefore, may not request evaluation resources from government grants, philanthropy, and other funders.

In addition, government and private funding sources often do not support evaluations, and those that do rarely require rigorous research designs that measure the causal impact of the program on key outcomes. Most foundations provide funding for evaluations for fewer than 10% of their grants (Beer, 2016). Government and private grants that require evaluations typically ask grantees to track the outputs supported by the grant or to measure some basic outcomes for those who receive services. Even when there are 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 non-profit 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, but many of these did not have rigorous research designs (Haskins, 2014, p. 165).

In many instances, a rigorous research design is simply not feasible because of capacity constraints—the program only serves a small number of people. Small sample sizes limit the power of statistical tests of the impact of the program, making it difficult to measure precisely the effect on key outcomes, even in cases where the true effect is large. In other instances, the program may serve a large number of clients, but an impact evaluation is not feasible because demand for services is not sufficient to generate an appropriate comparison group. For example, if a job training program serves all eligible applicants, this leaves no potential clients for the control group. Programs without excess demand for services among an eligible population can still measure program impact through alternative, quasi-experimental approaches that compare eligible clients to ineligible clients, but some of these approaches require even larger sample sizes, and providers are often not aware of these options.

Even when funding is available to support reliable evidence-building, social service providers may not collect or have access to the data necessary to evaluate outcomes. Many providers collect information about the clients they serve, tracking outputs such as the number of meals provided or number of beds filled in a homeless shelter. But collecting information on outcomes—the impact that the program had on the client rather than the services provided—is much less common. Outcome data can be more difficult to collect, particularly on intermediate and long-term outcomes achieved after a client has completed a program. In addition, few service providers have access to data on outcomes for a comparison group of individuals, which is necessary to isolate program impact.

Collecting outcome data in an evaluation can be an expensive proposition. Surveys can be used to collect information on outcomes, but an hour-long survey can cost upwards of $500. Fortunately, in many instances, administrative records already contain information on key outcomes such as employment, earnings, college persistence and completion, contact with the criminal justice system, and hospital admissions, among others. Moving to Opportunity and

other large-scale impact evaluations have relied heavily on these kinds of administrative data. The challenge is that these data are typically not available to social service providers and their research partners for impact evaluation purposes.

IV. Promoting More Evidence-Based Programming

As noted above, we do not see a stronger base of reliable evidence among human service nonprofits because rigorous impact evaluations often do not meet the short-term needs of these providers and because there are several practical barriers that make implementation of such evaluations challenging. Addressing these obstacles would help promote greater evidence of program impact among these providers.

If rigorous impact evaluations are to become more commonplace, evaluators will need to find ways to ensure that they fit the needs of the provider. This means more timely evidence that providers can more easily use to improve their program rather than simply to evaluate overall impact (Cody and Asher, 2014). Better, more accessible data will go a long ways towards making useful, near-real-time evidence possible.

To encourage more providers to embrace evidence-based programming, an impact evaluation needs to be characterized not as a high-stakes assessment but rather as an important diagnostic tool for determining how to continuously improve programming (Haskins, 2014, p. 234). This can be challenging when impact evaluations are conducted on programs that are already at scale, as the intervention model is often well-established. However, for new, smaller-scale programs, impact evaluations can provide critical information on ways to improve effectiveness. With a broad culture of evaluation, where interventions are continuously evaluated and this evidence leads to improved programming and the allocation of resources towards the most effective programs, the weakest programs will be less likely to reach the point where they are evaluated at scale. Thus, measuring impact for new programs will make it more likely that only the best programs are tested on a large scale.

Some of the practical barriers to conducting impact evaluations will be hard to break down. For example, it will always be challenging to measure the impact of small programs and those with little excess demand. The most fruitful way to promote impact evaluations for these programs would be to encourage the practice of constantly building evidence of promise through non-experimental means—for example comparing outcomes for clients served to those for a comparison group that is constructed to match the demographic characteristics of those who receive services (Blundell and Costa Dias, 2000). This evidence of promise can serve as the catalyst to raise the resources needed to scale up small programs to a size that is sufficient for a fully-powered experimental evaluation. In some cases, programs that lack excess demand can increase interest in the program by promoting it more broadly, or by expanding eligibility. Alternatively, these programs may measure the impact through quasi-experimental methods.[11]

[11] For example, regression discontinuity research designs are often used that compares outcomes for groups on either side of an eligibility threshold. For example, see Angrist and Lavy (1999), Jacob and L. Lefgren (2006), Ludwig and Miller (2007).

To address the issue of evaluation costs, governments and private charities need to do more to support rigorous evaluation of promising programs. For the public sector, this funding could be provided through either government agency initiatives and/or new legislation. A current example of the federally funded approach 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. In this 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 provide support for promising initiatives that currently lack rigorous evidence. These grants create a pipeline for innovative programs that, if proven effective, can be scaled up for broader impact. The evidence requirement for the top tier ("Scale-up") i3 grants includes one or more well-designed and implemented RCTs or quasi-experimental studies.

Occasionally, funding to test new programs is made available through legislation. 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 SNAP participants. Each of these pilot projects is required to have an independent evaluation that compares outcomes for households participating in the pilot to a "control group" of households not participating in the pilot. The legislation also requires participating states to make administrative data available in order to track outcomes. Pilot initiatives like this go a long ways in promoting better evidence based programming.

As noted above, service providers and their research partners often lack access to data on outcomes even though such data is available through administrative data sources. There would be much greater evidence on program impact if providers could link micro-level data on program participants (and a comparison group) to administrative data. Access to these data will not promote evidence-based programming, however, if providers are not prepared to use them. Simple changes in standard practices can ensure providers are ready. For example, most providers do not have information on an appropriate comparison group. By routinely collecting information on program applicants who are eligible but not served, and whose outcomes they can track over time, providers will be better able to measure program impact in a fairly rigorous way in both the short-run and long-run. Ideally, service providers would have clear and consistent protocols for how to link their data to administrative data. These protocols would include a standardized list of personal identifying information (i.e. name, date of birth, etc.) to collect that would allow them to link these data sources. Providers would also benefit from a standardized process for addressing privacy concerns when linking and sharing data, including clear and consistent protocols for informed consent.

Examples of administrative datasets that would be particularly useful to social service providers include: earning records, utilization of public benefits, hospitalization and other health outcomes, health care utilization, arrest records and other criminal justice-related information, credit reports, and education records. Making these data available to providers and their research partners would facilitate numerous impact evaluations, incentivize more researchers to focus on policy-relevant studies, and provide policymakers with better evidence of program impact and effectiveness—all resulting in the design of more effective social service programs and policies and, in turn, improved and accelerated outcomes.

Promoting evidence-based programming requires more than just producing strong evidence of effective programs. Other providers need to act on this evidence so that the best programs are scaled up and replicated. Many social service providers looking to design new programs or improve existing programs do not have ready access to information about the most effective programs. Even when such information is available, it can be difficult for providers to sift through numerous studies and/or to separate strong evidence from weak evidence. In addition, there is often limited information about how to successfully implement and continuously evaluate programs with evidence of effectiveness.

In order to design and implement evidence-based programs, social service providers need a way to easily track down and navigate the existing body of 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 in the social services sector. There are already several clearinghouses with important information about effective programs, which can serve as a foundation for moving forward.[12] One important challenge here is helping stakeholders distinguish among the different tiers of evidence. This means that stakeholders need clear standards for what constitutes reliable evidence. Ideally, an independent entity would assess evaluations and identify those that are reliable.

However, reliable evidence alone is not enough to ensure that effective programs are implemented broadly—because providers rarely have access to information on how to successfully replicate these proven programs. Thus, reliable evidence should be complemented by 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 from the initial program.

V. Conclusion

A critical first step to ensuring that government programs are backed by credible and reliable evidence is to understand what works best for human service nonprofits. While there is still relatively little reliable evidence on the effectiveness of social service programs, some of the barriers to greater evidence are surmountable. Policymakers could accelerate the pace and quality of evidence-building by providing more resources for reliable impact evaluations, streamlining and standardizing access to key administrative data, and expanding support for the replication of effective programs. These efforts would significantly expand the prevalence of evidence-based programs at all levels of government so that scarce resources are allocated towards the most effective programs.

---

[12] A well-designed model of how to synthesize a large body of evidence is the What Works Clearinghouse, which is run by the U.S. Department of Education's research arm: the Institute of Education Sciences (IES). The U.S. Department of Labor offers a similar service for labor topics through the Clearinghouse for Labor Evaluation and Research (CLEAR). Outside the government, the Coalition for Evidence-Based Policy provides a nice one stop shop for what works in social policy. In 2015, the Coalition was integrated into the Laura and John Arnold Foundation. Its key content will soon be migrated to http://www.arnoldfoundation.org/initiative/evidence-based-policy-innovation/, and will be regularly updated on that site.

VI. References

Angrist, J., and V. 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 (2010), "Federal Programs for Youth: More of the Same Won't Work," Brookings (https://www.brookings.edu/opinions/federal-programs-for-youth-more-of-the-same-wont-work/).

Beer, Tanya (2016), "Evaluation Demand and Capacity in the Social Sector," presentation before the Commission for Evidence-Based Policymaking, November, 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*, vol. 21, no. 4, pp. 427–468.

Bridgeland, John and Peter Orszag (2013), "Can Government Play Moneyball?" The Atlantic, July/August. http://www.theatlantic.com/magazine/archive/2013/07/can-government-play-moneyball/309389/.

Cima, Rosie (2016). DARE: The Anti-Drug Program That Never Actually Worked. Priceonomics. December, 19 2016.

Cody, Scott and Andrew Asher (2014), "Smarter, Better, Faster: The Potential for Predictive Analytics and Rapid-Cycle Evaluation to Improve Program Development and Outcomes," in Policies to Address Poverty in America, Melissa Kearney and Ben Harris eds. Hamilton Project, June, 147-155.

Ennett, ST, Tobler, NS, Rigwalt, CL and RL Flewelling (September 1994). How effective is drug abuse resistance education? A meta analysis of Project DARE outcome evaluations. American Journal of Public Health. 1994;84:1394–1401.

GAO (2003). Youth Illicit Drug Use Prevention: DARE Long-Term Evaluations and Federal Efforts to Identify Effective Programs. GAO-03-172R: Published: Jan 15, 2003. Publicly Released: Jan 15, 2003.

Haskins, R., & Margolis, G. (2014). Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy. Brookings Institute Press.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. Journal of Public Economics, 94(1), 114-128.

Jacob, B. A., and L. Lefgren, L. 2006. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics*, 86 (1): 226-244.

Ludwig, J., and D. Miller. 2007. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, *Quarterly Journal of Economics*, February, 159-208.

Manzi, J. (2012). *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York: Basic Books.

Nussle, Jim and Peter Orszag (2014). "Let's Play Moneyball," in Moneyball for Government. J. Nussle, & P. Orszag, eds. Disruption Books.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). Lifetime effects: The High/Scope Perry Preschool study through age 40. HighScope Press.

Surgeon General (2001). Youth Violence: A Report of the Surgeon General. Chapter 5: Prevention and Intervention. National Center for Injury Prevention and Control (US); National Institute of Mental Health (US); Center for Mental Health Services (US). Rockville (MD): Office of the Surgeon General (US).

U. S. Department of Education (2003), Planning and Evaluation Service, Elementary and Secondary Education Division, "Third National Even Start Evaluation: Program Impacts and Implications for Improvement," Washington, D.C., http://www2.ed.gov/rschstat/eval/disadv/evenstartthird/toc.html.

West, Steven and Keri O'Neal (2004). Project D.A.R.E. Outcome Effectiveness Revisited. Am J Public Health. 2004 June; 94(6): 1027–1029. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1448384/