ChangeThis

250.03

MISMEASURING IMPACT:
THE GOLD STANDARD MOVEMENT'S THREAT TO THE NONPROFIT SECTOR

Nicole P. Marwell & Jennifer E. Mosley

The problems that social programs are intended to solve—such as poverty, poor health, crime, and mental illness—have been with us for millennia, and never seem to go away.

This means that efforts to eliminate or reduce those problems often are viewed with suspicion, and we have regular cycles of skepticism regarding the benefits of social programs. In the United States, where most such programs are delivered through a partnership between government and nonprofits, this skepticism affects both sectors.[1] Government–federal, state, and local–supplies much of the money to support social programs, but it often contracts with nonprofits to deliver these programs to people in their communities. Between the chronic nature of social problems, the skepticism regarding social programs' ability to solve them, and the ongoing shift in governmental assistance to the needy from cash support to service-based support delivered through nonprofits,[2] the pressure for these organizations and their government funders to prove the value of social programs keeps growing.

It is within this larger context that we see the rising interest in using randomized controlled trials (RCTs) to evaluate nonprofits. Indeed, the promise that RCTs can deliver scientific clarity about which social programs work has been widely accepted. And yet, we learned in our research that many nonprofit sector professionals with RCT experience have deep reservations about the ability of the method to deliver on that promise. This left us with a puzzle: if RCTs in fact mostly fall short in helping nonprofits meet important evaluation challenges, why is it now commonplace to claim that RCTs are the "gold standard" for evaluating nonprofits?

RCT did not find its place at the top of the evidence hierarchy simply on its own merits …

THE EVIDENCE BATTLE

Many people concerned with making sure social programs–and the nonprofits that deliver them–are improving people's lives have embraced the idea of a "hierarchy of evidence." This formulation suggests that there are better and worse types of evidence, and that the RCT naturally sits atop the hierarchy. But the RCT did not find its place at the top of the evidence hierarchy simply on its own merits: this placement was constructed through what we call the "Gold Standard movement."

The simplest story of the Gold Standard movement goes like this. Sometime around 1980, a growing number of economists became dissatisfied with the then-dominant approach to doing microeconomics, leading to what has been called the "credibility revolution."[3] At the time, most microeconomic research that sought to inform public policy decisions relied on building econometric models from theory, then testing the models with observational data. Critics argued that results were highly contingent on a model's underlying theoretical assumptions; quite different results occurred when different assumptions were used.[4] Credibility revolution scholars argued that if we wanted to determine whether the changes observed in social program participants were in fact the result of participating in that program, experimental (that is, RCT) research designs would be required.[5]

Nailing down whether a program is causing change in its participants is a hard question because of the counterfactual. When someone takes part in, for example, a job training program, we can only observe what happens to them afterwards: if they got a job, what kind of job, at what wages, and so on. We cannot also observe the counterfactual: what would have happened to them in terms of employment if they had not participated in the program.

By 2010, two economists active in the credibility revolution would write that RCTs had delivered some of the "most influential microeconometric studies to appear in recent years," providing "results that are defensible both in the seminar room and in a legislative hearing."[6] And in 2019, the Prize in Economic Sciences in Memory of Alfred Nobel was won by three of the most high-profile practitioners and promoters of RCTs, in recognition of how their approach had upended the status quo in international development economics' search to alleviate global poverty. This simple version of the Gold Standard movement's history tells us that RCTs rose to prominence simply because they provide the best evidence for understanding whether or not a social program works.

Scholars who have delved into the history of economics during this period, however, offer a second version of the story, one that is decidedly more complex.[7] In this version, significant points of contention have always existed regarding the reliability and validity of RCT evidence—notwithstanding the "credibility revolution." Indeed, objections to the idea that RCTs necessarily offer superior evidence have been ongoing in multiple fields, including economics.[8] Some economists and other social scientists have also argued that RCTs are unethical because they deny people access to a potentially helpful program simply to facilitate scientific investigation into the program's effects.[9] In addition, broader challenges affecting scientific progress also apply to RCTs, such as the prevalence of "p-hacking"[10] (searching for statistically significant findings in data analysis rather than testing theoretically informed hypotheses), the lack of publication of null findings[11] (which biases evidence in favor of a hypothesis by limiting the availability of evidence that does not support it), and allegations of data falsification.[12]

Alongside the disagreements among economists about whether RCTs should be considered the pinnacle of an evidence hierarchy, researchers specializing in policy and program evaluation have also weighed in on the evidence battle. A primary concern of evaluation scholars is that the research method chosen be a good match for the evaluation question at hand.[13] The RCT, however, is well-suited only to one type of question: Did x program cause a change in y (usually narrowly specified) outcome? Nonprofit organizations may at times have an interest in such a question, but they also have many other important questions to which they seek answers, such as whether community members can access their programs, or if the program model they are using takes account of the particular needs of their target population. So what is the state of the evidence today? While RCTs certainly have strong scientific advocates, there is far more discontent with the method than proponents of the Gold Standard movement let on.[14] Indeed, the construction of the hierarchy of evidence on which the Gold Standard movement relies has been at least as much of a social process as a scientific one.[15]

THE FUNDING BATTLE

Understanding why some types of evidence carry more authority than others requires analysis of the context in which evidence is being deployed. The struggle over what kinds of evidence government should rely on when making decisions about spending on social programs is tightly tied to the skepticism that those programs—jointly provided by government and non-profits—regularly face about their value. The same could be said about the role of government spending overall, with one of the fundamental political disagreements in the nation being about whether such spending should be expanded or curtailed. As the sociologist Elizabeth Popp Berman recounts, between about 1950 and 1980, a strategy for providing better answers to this question—in the form of finding out whether government spending was achieving its articulated goals—was developing inside the federal government. This strategy was grounded in "economic thinking,"[16] an approach in which effectiveness and efficiency took center stage in determining what policies government should pursue.

The construction of the hierarchy of evidence on which the Gold Standard movement relies has been at least as much of a social process as a scientific one.

For example, what is the most effective way for government to address the needs of people who are poor? One of the first social policy RCTs asked a version of this question to assess how much a negative income tax would reduce the nation's poverty rate.[17] Another early social policy RCT sought to understand whether requiring patients to share the costs of their government-sponsored health insurance would affect how much health care they used.[18] These early RCTs paved the way for asserting the importance of causal evidence to assess government-supported social programs. They also helped to build an entire industry of professional evaluation organizations, which were needed to provide quick answers to policymakers' questions about what sorts of causal effects different policies might produce. [19] The growth of this industry was helped along by fast-rising allocations of federal funds: by 1968, each time a new program received federal funding, one percent of its cost was allocated to the evaluation of its results.[20] These early stirrings of the Gold Standard movement—the label we use to refer to organized efforts to promote RCTs and causal evidence in public policymaking and the nonprofit sector—thus married the pursuit of causal evidence with funding opportunities.

Over the next several decades, this was the way interested parties put the building blocks of the Gold Standard movement into place, taking steps to ensure causal evidence would play an increasingly important role in policymaking. Members of that movement refer to their work as advancing "evidence-based policy."[21] This is misleading, however, because many scholars and practitioners outside the Gold Standard movement agree that policy should be evidence-based—they just advocate for a wider range of evidence to be considered.[22] Still, the most concerted and powerful efforts to advance the use of evidence in policymaking have focused specifically on causal evidence, which comes only from RCTs and (less desirably) quasi-experimental methods.

SPREADING RCTS TO U.S. NONPROFITS: THE SOCIAL INNOVATION FUND

Researchers pioneering RCTs in international development often collaborated with international nongovernmental organizations (NGOs) to test whether particular social programs were effective. Working with NGOs offered some distance from concerns about democratic governance and the proper role of the state in an RCT that, for example, withheld state-sponsored services from the control group.[23] The experience of these researchers offered guidance for later efforts to conduct RCTs inside U.S. nonprofit organizations; indeed, the Abdul Lateef Jamal Poverty Action Lab (J-PAL), founded by two of the 2019 winners of the Prize in Economic Sciences in Memory of Alfred Nobel, now has a robust set of U.S.-based RCTs. Many of these are being conducted in partnership with nonprofit organizations.[24]

The vision laid out by these researchers was compelling to the data-driven Obama administration. which worked to elevate the importance of causal evidence in the development and funding of government social programs.[25] This effort included the first systematic effort to get U.S. nonprofits to subject their programs to rigorous evaluation: the Social Innovation Fund (SIF).[26] Between 2010 and 2016, the SIF made hundreds of millions of dollars in grants to thirty-nine intermediary organizations—nonprofits whose principal work is funding or supporting service-providing nonprofits—which in turn made sub-grants to just under three hundred nonprofits that were operating promising programs in local communities across the country.[27] Built into the grants to support these nonprofits' program work was a requirement that they undertake rigorous evaluation—generally, RCT or quasi-experimental evaluation—of their program impacts.

But the SIF evaluation experience underlines how challenging it is for nonprofits to conduct RCTs, or even quasi-experimental evaluations. Indeed, while a 2016 report on the SIF indicates that the initiative had some three hundred sub-grantees,[28] only around eighty evaluations actually were completed.[29] Of these eighty or so evaluations, only thirty-two assessed program outcomes or impacts, and only half of those thirty-two were adequately powered (that is, had a large enough sample size in both comparison groups) to provide credible evidence on at least one outcome.[30] To sum up: three hundred nonprofit organizations were asked by the SIF to conduct a high-quality evaluation study, and only sixteen of them delivered.

Many scholars and practitioners outside the Gold Standard movement agree that policy should be evidence-based—they just advocate for a wider range of evidence to be considered.

CHANGING THE CONVERSATION

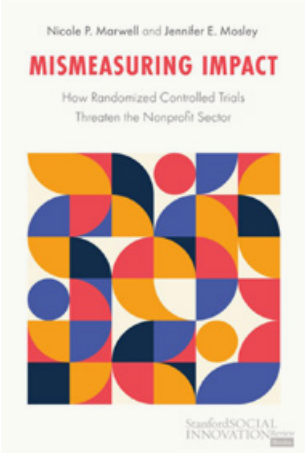
The SIF evaluation experience offered an early sign that RCTs are a poor match to evaluate the complex activities of nonprofit organizations.[31] Nevertheless, many nonprofit sector stakeholders feel compelled to discuss and advocate for the use of RCTs to evaluate nonprofit programs and organizations. The success of the Gold Standard movement in the funding battle has been critical to this development—especially its efforts to increasingly tie government funding for nonprofits to the use of programs with RCT evidence of effectiveness. This has been occurring despite the ongoing evidence battle over whether RCTs of social programs actually deliver the scientific results their advocates claim they do.

In *Mismeasuring Impact: How Randomized Controlled Trials Threaten the Nonprofit Sector*, we draw on our own research to flesh out the five problems with using RCTs in nonprofits. Our evidence comes from interviews with professionals in the nonprofit sector—nonprofit managers, professional evaluators, philanthropic foundation program officers—who have experienced first-hand the growing use of RCTs in the sector. The problems these professionals helped us identify touch on the evidence battle, the funding battle, and broader questions of the role of nonprofits in society. **We learned that there are important limits to using RCTs to evaluate nonprofits and—like the experts in nonprofit evaluation on whose work we draw—we want to change the conversation about how to use evaluation to more fully meet the needs of nonprofit organizations and the communities they serve.**

Mismeasuring Impact

Nicole P. Marwell & Jennifer E. Mosley

Info



Ready to dig deeper into the book?
Buy a copy of Mismeasuring Impact.

Want copies for your organization or for an event?
We can help: customerservice@porchlightbooks.com
800-236-7323

ABOUT THE AUTHORS

Nicole P. Marwell and Jennifer E. Mosley are Professors at the Crown Family School of Social Work, Policy, and Practice at the University of Chicago. Their research on nonprofit organizations has been published widely in leading journals in the fields of nonprofit studies, sociology, public administration, and social work.

SHARE THIS

Pass along a copy of this manifesto to others.

SUBSCRIBE

Sign up for e-news to learn when our latest manifestos are available.

Porchlight

Curated and edited by the people of Porchlight, ChangeThis is a vehicle for big ideas to spread. Keep up with the latest book releases and ideas at porchlightbooks.com.

This document was created on July 16, 2025 and is based on the best information available at that time.

Endnotes

1. Salamon, Partners in Public Service; Steven Rathgeb Smith and Michael Lipsky, *Nonprofits for Hire: The Welfare State in the Age of Contracting*.
2. Scott W. Allard, *Out of Reach: Place, Poverty, and the New American Welfare State*; Nicole P. Marwell, "Privatizing the Welfare State."
3. Angrist and Pischke, "The Credibility Revolution in Empirical Economics."
4. Edward E. Leamer, "Let's Take the Con Out of Econometrics"; LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data"; Fraker and Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs."
5. Orley Ashenfelter, "The Case for Evaluating Training Programs with Randomized Trials"; LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data."
6. Angrist and Pischke, "The Credibility Revolution in Empirical Economics," 4.
7. This discussion is indebted to the analyses provided in Daniel Breslau, In Search of the Unequivocal and Elizabeth Popp Berman, *Thinking Like an Economist*.
8. Heckman, "Randomization and Social Policy Evaluation"; Angus Deaton, "Instruments, Randomization, and Learning About Development"; Deaton, "Evidence-Based Aid Must Not Become the Latest in a Long String of Development Fads"; Lant Pritchett and Justin Sandefur, "Context Matters for Size"; Deaton and Cartwright, "Understanding and Misunderstanding Randomized Controlled Trials"; Heckman, "Randomization and Social Policy Evaluation Revisited."
9. See, for example, Timothy Ogden, "Lant Pritchett"; Phyllis Solomon, Mary M. Cavanaugh, and Jeffrey Draine, "Ethical Considerations of Randomized Controlled Trials"; Pritchett, "Randomizing Development"; Florent Bédécarrats, Isabelle Guérin, and François Roubaud, eds., *Randomized Control Trials in the Field of Development*; Bédécarrats, Guérin, and Roubaud, "All That Glitters Is Not Gold."; For a robust summary of the arguments on how to conduct ethical RCTs, see Feeney et al., "Ethical Conduct of Randomized Evaluations."
10. John P. A. Ioannidis, "Why Most Published Research Findings Are False"; Robert J. MacCoun, "P-Hacking: A Strategic Analysis."
11. Annie Franco, Neil Malhotra, and Gabor Simonovits, "Social Science. Publication Bias in the Social Sciences."
12. See, for example, the website "Data Colada," one of a number of online sources devoted to identifying publications with falsified data.
13. Benjamin, Ebrahim, and Gugerty, "Nonprofit Organizations and the Evaluation of Social Impact"; Mary Kay Gugerty and Dean Karlan, The Goldilocks Challenge; Eleanor Chelimsky, "Factors Influencing the Choice of Methods in Federal Evaluation Practice"; Carol H. Weiss, *Evaluation: Methods for Studying Programs and Policies*
14. Malte Neuwinger, "What Makes Randomized Controlled Trials so Successful—for Now?"; Jennifer E. Mosley et al., "Impact, Equity and Philanthropic Foundations."
15. Breslau, *In Search of the Unequivocal*
16. Berman, *Thinking Like an Economist*.
17. David N. Kershaw, "A Negative-Income-Tax Experiment"; David N. Kershaw and Jerilyn Fair, eds., *The New Jersey Income-Maintenance Experiment*.
18. Joseph P. Newhouse and Insurance Experiment Group, "Free for All?: Lessons from the RAND Health Insurance Experiment."
19. Daniel Breslau, "Contract Shop Epistemology"; Breslau, *In Search of the Unequivocal*; Berman, *Thinking Like an Economist*; Alice O'Connor, *Poverty Knowledge*.
20. Berman, *Thinking Like an Economist*.
21. For example, Jon Baron, "A Brief History of Evidence-Based Policy"; Haskins and Margolis, *Show Me the Evidence*.
22. Kathryn Oliver and Annette Boaz, "Transforming Evidence for Policy and Practice"; Justin O. Parkhurst and Sudeepa Abeysinghe, "What Constitutes 'Good' Evidence for Public Health and Social Policy-Making?"; Andrew D Oxman, "The Cochrane Collaboration in the 21st Century."
23. Luciana de Souza Leão, "What's on Trial?"; Ogden, "Lant Pritchett."
24. See the J-PAL database of evaluations, searchable by country, at https://www.povertyactionlab.org/evaluations.
25. Haskins and Margolis, *Show Me the Evidence*; Haskins and Baron, "Building the Connection Between Policy and Evidence."
26. Haskins and Margolis, *Show Me the Evidence*.
27. Xiaodong Zhang and Jing Sun, "Meta-Analysis of Evaluations Across the Social Innovation Fund Program"; Lily Zandniapour et al., "Strengthening Organizational Practice"; Lily Zandniapour and Mary Hyde, "Lessons from the Social Innovation Fund."
28. Zhang and Sun, "Meta-Analysis of Evaluations Across the Social Innovation Fund Program."
29. Zhang and Sun, "Meta-Analysis of Evaluations Across the Social Innovation Fund Program"; Scott Richman and Rebekah Selekman, "Scaling Evidence-Based Models: Document Review Rubrics"; Scott Richman and Andrei Streke, "Evidence of Effectiveness in AmeriCorps-Funded Interventions."
30. See page 18 of Zhang and Sun, "Meta-Analysis of Evaluations Across the Social Innovation Fund Program."
31. Benjamin, Ebrahim, and Gugerty, "Nonprofit Organizations and the Evaluation of Social Impact"; Lehn Benjamin, "Account Space: How Accountability Requirements Shape Nonprofit Practice"; Gugerty and Karlan, The Goldilocks Challenge; Ebrahim, Measuring Social Change; Julnes and Rog, "Editors' Notes"; George Julnes, "Editor's Notes: Experimental Methodology."