

BUILDING MORE HAYSTACKS, FINDING MORE NEEDLES

RYAN MARTIN

It is hearing day. For months, you have reviewed research, met with advocates, and called experts to get up to speed on what's happening.

A decade ago, Congress allocated tens of millions of dollars to start a new program focused on addressing a specific social problem. At the time, legislators—lacking evidence on what might work best—mandated the federal agency running this new program conduct an evaluation to see if it works.

Congress has waited a decade to know the answer, spent millions on studying the program, reviewed many interim reports, and now the final report is in. What did it find?

No significant impacts.

For the measures tracked by the study, there were some promising findings in the short term, but none of them lasted. It looks like there may have been some positive results for a subset of people in some places, but the report concludes with those four dreaded words: “more research is needed.”

You, your colleagues, and members of Congress have met with dozens of groups who are lobbying for the program over the last six months—sharing anecdotes of success, highlighting the many thoughtful organizations working hard to address the issue, even giving your boss an award for championing the cause. You have received letters with hundreds of signatures calling

for the extension and expansion of the program, seen op-eds placed in publications your boss reads, and heard how lobbyists at recent evening fundraisers reiterated the importance of this program to your boss. But does it work?

Today is the day Congress will decide. They are not voting on its fate, but the statements made at this hearing will set in motion a narrative that will harden as the program's profile rises and partisan viewpoints begin to take hold. Will key members of the committee support it and call for continued funding, hoping a few tweaks or another evaluation will show improvement? Or will they end it for good?

Millions of dollars are on the line, many reputations are at stake, and the fate of multiple nonprofits and dozens of local organizations hangs in the balance—not to mention the thousands of people participating in the program who hope it will make a real difference in their lives.

MOST THINGS WON'T WORK

I have seen this many times, and this scenario frequently plays out not only in Congress but in state legislatures and other decision-making bodies around the country.

Why? Addressing social problems is hard. There is no scientific formula or law of physics that says X action will cause Y reaction, guaranteeing less homelessness, fewer children in foster care, or higher earnings for those stuck in low-wage jobs. In fact, the only law related to the impact of a social program is that it is not likely to have any impact at all.

Known as *The Iron Law of Evaluation*, sociologist Peter H. Rossi argued in a 1972 paper¹—updated and spread widely in the late 1980s—that “the expected value of any net impact assessment of any large scale social program is zero.”² In making this claim, he notes “the Iron Law arises from the experience that few impact assessments of large scale social programs have found that the programs in question had any impact.” He noted there were exceptions and some programs had demonstrated positive results but that these were unfortunately few and far between. Speaking in 2003 about the impact of his earlier paper, he noted the Iron Law was, thankfully, not as iron-clad as it had first seemed, saying, “I believe that we are learning how properly to design and implement interventions that are effective.”³

The year 2003 was a long time ago. Is it really true that most social programs won't work? With all of our advances in understanding human behavior and the great leaps forward in technology and data analytics, haven't we cracked the code on how to change lives? Unfortunately, these efforts are still very much a work in progress.

In a July 2013 hearing held by the House Committee on Ways and Means, one expert noted how few studies of social programs had shown positive results. Summarizing the impacts of ninety randomized controlled trials in education, he noted about 90 percent found weak or no positive effects.⁴ The same thing was found with employment and training programs—75 percent of thirteen programs that had been rigorously evaluated showed little or no positive impact.⁵ In some cases, this has led to improvements in how evaluations are conducted to look more carefully for results. Unfortunately, however, this disappointment in finding that programs don't work sometimes means there is pressure to remove or weaken evaluation requirements, or to scale back evaluations because of the time and cost needed to conduct them (although easier access to administrative data is both speeding up and lowering the cost of such evaluations).

INEFFECTIVE PROGRAMS AREN'T HARMLESS

But remember, treating without testing can produce real harm. A program funded for a decade with the intention of helping low-wage workers move up the economic ladder is not benign if it doesn't work. Those who participated likely passed up other opportunities. They could have pursued a different education or training path, taken a new job, or even just stayed with the one they already had. They could have even ended up worse off than if they had never participated—in a lower-paying job or with less income, or having taken time away from work for a training that wasn't helpful. But even if the effects were not harmful to the individual themselves, there is still a large opportunity cost—what could have been done with those funds to truly help those needing a leg up, instead of spending on something that did not get them where they needed to go.

Is the answer, then, to avoid rigorously evaluating social programs? Or, since many things won't work, should funding be ended until we find the right answer? No. Instead, we need to fail with more certainty, more frequently, more cheaply, and much faster than ever before.

MORE HAYSTACKS = MORE NEEDLES

For some reason, the failure of a social program has been assigned an out-sized burden of shame, ridicule, and finger pointing compared with failure in other disciplines. The witness at the 2013 Ways and Means hearing cited above pointed this out, trying to put the failure of so many social programs in context. As he noted at the time—and as the world has witnessed first-hand during the COVID-19 pandemic—the majority of medications under development turn out not to work, with a large share not yielding positive impacts in larger studies even when initial findings look promising.⁶ In the business world, failures may be even more common, with thousands of studies conducted by Google and Microsoft on new products or strategies showing no significant effects.⁷

We need to get over it. Most things won't work, and that is ok. Given this, we have to think and work differently.

If we start from that point of view, we ask very different questions. If there are only a few needles in the haystack, how do we most quickly and effectively find them? Do we create one monolithic pile, take a sample, and say “on average, there were zero needles in this hay?” Do we create many different piles of hay but examine only one closely? No. We need lots of different haystacks, good detectives to look through them, and we need to do it over and over again.

As the former head of the Institute of Education Sciences put it,⁸ “the probability of finding [an effective program] will be remote unless we search widely, frequently, and intelligently. In short, experiment, experiment, experiment.”

We need to be all about finding the needles.

MAKING FAILURE A SUCCESS

If we know most social programs won't work, how can we make it so that finding something that doesn't work is acceptable and is seen as progress? Where program operators, funders, and evaluators aren't afraid to share results? In other words, how do we make failure a success? By treating the development of social programs the same way as we treat other disciplines.

The FDA approves the trial of a new drug to treat breast cancer. After promising results in an early study, the research is halted when a larger trial reveals no significant positive impacts and major side effects. Congress holds

a hearing on the failure of the drug, noting research showed it didn't work but also highlighting anecdotes from some who seemed to benefit. The hearing ends with members at odds on whether the drug works or not, and a few months later funding to develop breast cancer treatments is eliminated.

No, of course not.

Yet this is how the world works for social programs. Congress and other entities provide funding for a specific program, review how it worked, then decide on whether it should continue, be changed, or end. This not only makes the learning process extremely slow; it also makes evaluation incredibly high-stakes. These factors create an environment where there are competing pressures to continue or discontinue programs for reasons other than their effectiveness.

There has got to be a better way, and there is.

LEGISLATE THE PRIORITY, NOT THE PROGRAM

In health care, Congress does not dictate which treatments are allowed or which medicines are approved. Instead, they created a process specifying the *priority* and not the *product*. This same approach can be taken with social programs.

At the federal level, progress has been made in recent years to do just that—specify what the goal is while leaving the selection of the specific program to the state or local entity. This also allows programs to change over time as new evidence is developed. Examples include the following:

- The Every Student Succeeds Act (P.L. 114-95) provides local education agencies with flexibility to select programs that best meet their needs, with programs being ranked in one of four levels as having strong, moderate, or promising evidence, or as demonstrating a good rationale for expecting positive impacts.⁹
- The Maternal, Infant, and Early Childhood Home Visiting program (MIECHV—section 511 of the Social Security Act) requires funding be spent to achieve outcomes listed in the law, using programs that meet evidence requirements delineated by the U.S. Department of Health and Human Services (HHS). States can also use up to 25 percent of their funds on programs that have not yet met the evidence requirements but that will undergo a rigorous evaluation.¹⁰

- The Family First Prevention Services Act (P.L. 115-123) directed HHS to review programs for mental health, substance abuse, and parenting focused on preventing children from entering foster care. Programs are rated as promising, supported, or well supported based on a rigorous review of evidence.¹¹
- The Social Impact Partnerships to Pay for Results Act (P.L. 115-123) created a \$100 million fund to tie payment directly to outcomes. In this case, instead of creating a list of evidence-based programs, an entity can sign an agreement to receive payment only if they produce the desired social outcome.¹² Similar language focused on tying funding to outcomes also has been included in the Workforce Innovation and Opportunity Act (P.L. 113-128), the Every Student Succeeds Act (P.L. 114-95), the Bipartisan Budget Act (social impact partnerships demonstration projects and the Maternal, Infant, and Early Childhood Home Visiting program [P.L. 115-123]), and the Carl D. Perkins Career and Technical Education Act (P.L. 115-224).

The next step in this effort is to not just evaluate programs to see if they achieve the priority but to help providers produce better outcomes over time. In many cases, program operators lack the resources—whether that be time, expertise, money, or data—to analyze what is working and what is not and improve their practices as a result. For example, many small programs or those serving a particularly disadvantaged group may not show positive results in an evaluation yet they may also lack opportunities to further learn from the evaluation and try something different. Funding continuous improvement can help overcome this challenge by increasing the number of interventions that can achieve the goal instead of relying solely on evaluations identifying programs that are already working.

By making this approach more commonplace—and by supporting continuous improvement—Congress and others can speed up the development of more effective programs, as well as redirect funding toward approaches that yield the best results. These efforts also can create a climate where failure is acceptable, evidence building is prioritized, and those running programs adapt based on what has been learned instead of fighting for the status quo.

In a world where Congress and others fund a priority and not a program, the failure of one intervention is no longer an existential threat with the

potential to end investment in the issue. It just means it is time to learn what went wrong, improve that approach, or try another one instead. Funding to address the social problem does not go away. While it still may be difficult to identify what works and adapt to see what might be more effective, the priority still exists and support to address it continues.

With “tiered evidence” designs where financial support is provided to programs with varying levels of evidence, funding is not for a specific program but, instead, directed toward interventions that address the priority. If those who do not demonstrate results at first also receive support, this structure can help them identify improvements so they can become more effective over time, so that even a “failed” program can try again to achieve the goal.

Tying funding to outcomes can work in a similar way, shifting spending toward successful programs as well as potentially speeding the development of new ideas. “Pay for performance” often relies on real-time measurement of results as well as an evaluation of longer-term outcomes, so those providing services have strong incentives to monitor progress and adapt as needed to ensure they achieve the goal. This approach also has shown an ability to draw private investment and business expertise into social programs, which often brings with it a level of analysis and performance management not traditionally available to social service providers—allowing them to innovate and improve as they go.

A POSSIBLE FUTURE

It is hearing day. For months, you have reviewed the research, met with advocates, and called experts to get up to speed on what’s happening.

A decade ago, Congress allocated tens of millions of dollars to address a social problem. At the time, legislators—lacking evidence on what might work best—mandated the federal agency running the new program help organizations build evidence of what works.

Congress has waited a decade to see what progress has been made, spent millions studying the impact, reviewed many interim reports, and now the final report is in. What did it find?

Progress. Specific programs (and certain features of other programs) have been shown to move the needle on improving people’s lives. Those that work have been replicated and expanded, and new programs are in the pipeline that look promising. There are lots of failures, but some successes pointing the way to designing better programs.

Congress decides to continue the funding, clarify the goals they want to achieve, and let the thousands of people across the country working to solve this problem keep innovating. The message is clear: Build more haystacks, and keep looking for more needles.

NOTES

1. See the Welfare Reform Academy page at University of Maryland School of Public Policy website, Rossi's Remarks Iron Law Revisited, http://www.welfareacademy.org/rossi/Rossi_Remarks_Iron_Law_Reconsidered.pdf.

2. Peter H. Rossi, "The Iron Law of Evaluation and Other Metallic Rules," See this 1987 paper at www.gwern.net/docs/sociology/1987-rossi.pdf.

3. In Peter Rossi, "The 'Iron Law of Evaluation' Reconsidered," presented at 2003 AAPAM Research Conference, Washington DC., October 2003, www.welfareacademy.org/rossi/Rossi_Remarks_Iron_Law_Reconsidered.pdf.

4. Statement of Jon Baron, president of the Coalition for Evidence-Based Policy. House Committee on Ways and Means, Subcommittee on Human Resources Hearing on What Works, July 17, 2013, <http://coalition4evidence.org/wp-content/uploads/2013/07/Testimony-before-Ways-and-Means-HR-subcommittee-7.17.13-Jon-Baron.pdf>.

5. Ibid.

6. Ibid.

7. Ibid.

8. Russ Whitehurst, "Acceptance Remarks, November 9, 2007, www.welfareacademy.org/rossi/2007_whitehurst_speech.shtml.

9. See ESSA Tiers of Evidence: What You Need to Know page at REL Midwest website, <https://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/blogs/RELMW-ESSA-Tiers-Video-Handout-508.pdf>.

10. Emily Sama-Miller, Julieta Lugo-Gil, Jessica Harding, and others, "Home Visiting Evidence of Effectiveness (HomVEE) Systematic Review. Handbook of Procedures and Evidence Standards: Version 2," p. 2, December 2020, HomVEE, https://homvee.acf.hhs.gov/sites/default/files/2021-02/HomVEE_Final_V2_Handbook-508.pdf.

11. See Welcome page at HHS clearinghouse website, <https://prevention.services.abtsites.com>.

12. See Treasury page, <https://home.treasury.gov/services/social-impact-partnerships/sippra-pay-for-results>.