How This Works

REASONS FOR CAUTION AND OPTIMISM

TRULY MAKING A DIFFERENCE
The Outcome/Impact Dilemma

Gordon Berlin
MDRC

I was a young staffer at the U.S. Department of Labor years ago when the National Supported Work Demonstration findings were released.[1] I learned a critical lesson that has played out in multiple policy areas many times since: It is crucial to distinguish between *gross outcomes* (such as the percentage of program participants who enter a job or cycle off welfare) and *net impacts* (the improvement in these outcomes that was actually attributable to program participation). Failure to heed this lesson threatens to derail what can otherwise be a helpful shift now underway toward an outcome orientation to guide program improvement. Outcomes alone are often not a reliable metric for judging the effectiveness of social investments.

In the National Supported Work Demonstration, which served four subgroups of disadvantaged individuals, the data on gross outcomes and net impacts pointed in opposite directions. The most favorable net impacts on employment were found for long-term welfare recipients, the subgroup that had the *lowest* employment outcomes—meaning that the program worked best for the subgroup that had the worst absolute outcomes. Yet, there was no value added for the subgroup with the *best* employment outcomes, since a randomly assigned control group showed that its earnings would have increased just as much without the Supported Work services.

This and numerous subsequent evaluations underscore the reality that much of what is often credited as *program-produced* outcomes can be driven more by other factors, such as a strong economy or the natural progress that highly motivated participants selected for the program would have made over time anyway. Perhaps some of the confusion can be avoided if we're clear that the term "outcomes" does not apply only to

1 MDRC Board of Directors, "Summary and Findings of the National Supported Work Demonstration" (1980).

program participants; members of a control group (and everyone else who doesn't participate in the program) have outcomes, too.

The point is not that net impacts matter and gross outcomes do not; both measures are important, but they tell us different things that must be reconciled in the ongoing effort to build evidence on program effectiveness—and continuous improvement—under real-world conditions. As a practical matter, program operators must use gross outcomes to monitor performance and motivate staff; they typically don't have access to a control group to tell them how much their services have actually *caused* the outcomes they strive for and observe.[2] Yet, policy, funding, and program design decisions should be based more on net impacts to avoid pouring scarce resources into programs with high gross outcomes but limited, if any, value added. Excessive pressure to increase outcomes can backfire by skewing program providers' incentives, since the fastest way to show "improvement" is often to simply screen out harder-to-serve individuals—ironically, the very people who might benefit most from the services offered.

There are no easy solutions to this dilemma, but some important steps can help to identify potential distortions and at least partially overcome the risks. Other authors have explored some of the implications of the "outcome mindset,"[3] so I will elaborate briefly here on two aspects of the interplay between gross outcomes and net impacts: (1) how the characteristics of program participants can influence results; and (2) developing performance measures in the context of scaling and replicating programs.

UNDERSTANDING THE CHARACTERISTICS OF THE INDIVIDUALS BEING SERVED

It is particularly important to interpret program outcomes in light of participants' characteristics that are likely to be correlated with the intended outcomes: for example, when a job training program enrolls participants with strong employment histories, education levels, and motivation to work; or when a preschool program mostly serves children who come from stable families and were in a similar program the year before.

Program outcomes should also be interpreted in light of how participants were recruited and the extent to which the selection process may have screened out individuals who are more difficult to serve. This can be clarified by conducting a "funnel analysis," which documents key steps in the participant recruitment and selection process (such as eligibility criteria, interviews, or testing) and the reasons why individuals fail to come forward or drop off along the way, especially if it results from decisions that the program's staff makes. A tipoff that the selection process may be whittling down enrollees to the most motivated group: programs often start the enrollment process with many times the number of potential participants who ultimately enroll.

A focus on participants' characteristics might also enhance the prospects of producing favorable net impacts. For example, a recent data analysis from the National Head Start Impact Study showed that dual-language learners and Spanish-speaking children with low early literacy and math skills when they entered Head Start had gains that doubled the average net impact for the full sample.[4] Moreover, Head Start's impacts were largely driven by children who, in the absence of Head Start, would have stayed in home-based care throughout the day.[5] Targeted outreach to ensure inclusion of these and other underserved children might help boost the program's net impacts—although perhaps not the gross outcomes.

PERFORMANCE MEASURES IN THE CONTEXT OF SCALE-UP AND REPLICATION

The pressure for accountability and an outcome orientation goes hand in hand with growing interest in expanding cost-effective interventions. Since the stakes rise as more dollars are invested and more people are served,

2 Even in net impact studies, evaluators need to be careful to collect accurate information on the experiences of the control group. The actual "treatment difference" between the services that programs deliver to participants and the services that members of the control group receive elsewhere is sometimes less in practice than what had been assumed, leading to disappointing (and potentially misleading) findings.

3 Patrick Lester, "The Promise and Perils of an 'Outcome Mindset'," Leland Stanford Jr. University (2015); see also Michael Bangser, "A Funder's Guide to Using Evidence of Program Effectiveness in Scale-Up Decisions," MDRC and Social Impact Exchange (2014); and Judith Gueron, "Throwing Good Money After Bad: A Common Error Misleads Foundations and Policymakers," Leland Stanford Jr. University (2005).

4 Howard Bloom and Christina Weiland, "Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study," MDRC (2015).

5 Avi Feller et al., "Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings" (January 11, 2016, draft).

substantial expansion should be guided by evidence of net impact. Indeed, there are a number of examples in both the public and philanthropic sectors that use "tiered evidence" approaches, which stimulate new model development and program improvement. However, they require that early positive outcomes be followed by more rigorous evaluations to confirm that the programs are producing a positive net impact at key stages in the development and scaling process.[6]

Although every iteration of a program with positive net impacts cannot realistically be evaluated in another round of impact evaluations, the program outcomes can be judged by asking the following questions:

1. Are the program components that drove the positive results being implemented with the same intensity and quality as in the original net impact evaluation?
2. Is the program serving a population with the same challenges, or has it enrolled an easier-to-serve group?
3. Are the same outcome measures being used as in the original evaluation?
4. How does the operating context, such as the policy environment and local economy, compare with that of the original evaluation?

These questions suggest the need for caution in assuming that the encouraging net impacts found in the original evaluation will necessarily be repeated when programs are scaled up or replicated. For example, the population being served might shift, funding constraints often force providers to sacrifice certain elements of the original program model, and it can be difficult to maintain the same level of quality when programs operate on a much larger scale or in different settings.

WHERE DO WE GO FROM HERE?

The full potential of evidence-based policy will be realized only if program operators, policymakers, and evaluators join in using the right mix of data on program outcomes and net impacts. Reliance on the wrong measures, lack of data on key measures, and poor-quality data can lead to faulty conclusions. Since many program providers' data systems are underfunded and data collection can be a primary driver of an evaluation's costs, public and private funders will need to support data collection infrastructure. In return, evaluators must find ways to collect, analyze, and report on the data as expeditiously as possible.

A fertile opportunity is presented by a next generation of researcher-practitioner partnerships that leverage the strengths of both entities to refine interventions. For example, one such collaboration capitalizes on a readily available database to embed predictive analytics, random assignment, and other rigorous methods to test alternative strategies for continuous improvement in a network of schools. The initial focus has been on using early warning systems to identify students who are at risk of failing to graduate and real-time data from quick-turnaround random assignment studies to test the net impact of various approaches to boosting students' attendance.

Even with such advances, bridging the distinction between program outcomes and net impacts will no doubt continue to be a struggle for the field. But if we are vigilant for potential distortions and are open to exploring promising options, such as effective researcher-practitioner partnerships, we can use both outcome and net impact measures to produce the right kind of evidence to guide policy and practice. If we are not vigilant, two key constituents—participants and taxpayers—will be the first to realize that the numbers we are touting don't actually signal any net benefit at all.

GORDON BERLIN *is president of MDRC, a nonprofit, nonpartisan research organization based in New York City and Oakland, California, that is dedicated to learning what works to improve the lives of low-income people. Before joining MDRC in 1990, he was executive deputy administrator for management, budget, and policy at the New York City Human Resources Administration. He also was deputy director of the Ford Foundation's Urban Poverty program and worked as a program analyst and project officer in the U.S. Department of Labor's Employment and Training Administration. Berlin was the founding executive director of the Social Research and Demonstration Corporation, a Canadian nonprofit social policy research organization. Berlin has authored and co-authored*

6 The federal Office of Management and Budget has been a catalyst in requiring rigorous evidence to support program scale-up. Federal initiatives, such as the Social Innovation Fund and the Investing in Innovation (I3) Fund, have enlisted private partners to match public funding in an ongoing process of evidence-building. Philanthropy-led efforts, such as the Edna McConnell Clark Foundation's investment in youth-serving organizations, calibrate funding to the level of evidence while also investing in bringing the evidence to higher levels.

numerous publications, including Learning from Experience: A Guide to Social Impact Bond Investing; Poverty and Philanthropy: Strategies for Change; Rewarding the Work of Individuals: A Counterintuitive Approach to Reducing Poverty and Strengthening Families; *and* What Works in Welfare Reform: Evidence and Lessons to Guide TANF Reauthorization.