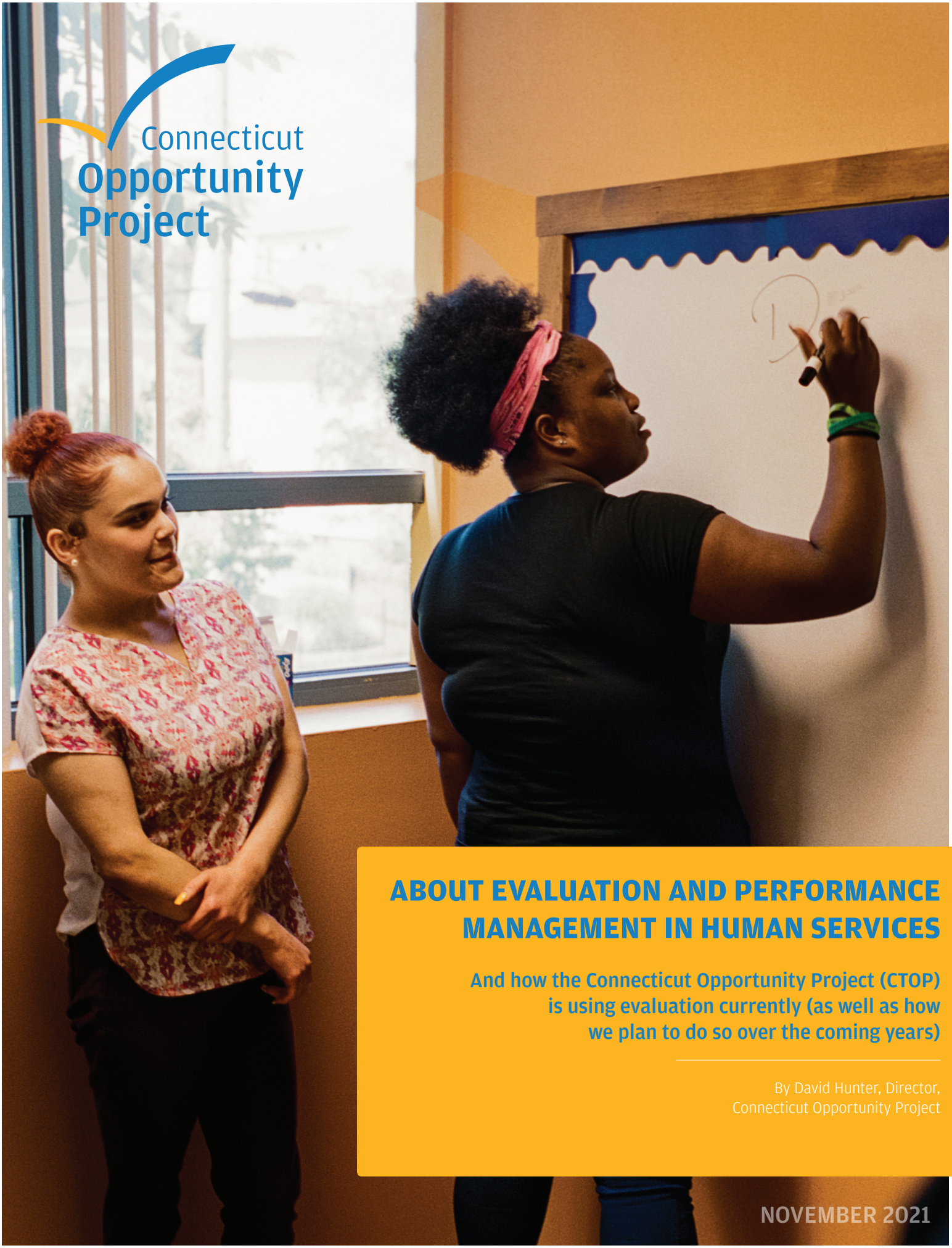


Connecticut Opportunity Project

ABOUT EVALUATION AND PERFORMANCE MANAGEMENT IN HUMAN SERVICES

And how the Connecticut Opportunity Project (CTOP) is using evaluation currently (as well as how we plan to do so over the coming years)

By David Hunter, Director,
Connecticut Opportunity Project

NOVEMBER 2021

Connecticut Opportunity Project

An initiative of Dalio Education, led by Barbara Dalio as part of Dalio Philanthropies, the mission of the Connecticut Opportunity Project is to invest in and help strengthen youth-serving organizations in Connecticut so they can work effectively, reliably, and sustainably with young people ages 14 to 22 who are at risk for dropping out of school (disengaged youth) or who already have done so (disconnected youth) – in order to help them re-engage in and complete secondary education and transition successfully to the pursuit of post-secondary education, such as a technical certification, military enlistment, or an academic degree – with the ultimate goal that they will achieve satisfying employment that supports their agency and self-sufficiency.

TABLE OF CONTENTS

> If the numbers you are looking at are boring . . . you are looking at the wrong numbers.

— Edward R. Tufte

Foreword: What Is Evaluation?

While there are many definitions of evaluation, there is general agreement that it consists of scientifically responsible methods for determining the merit of a policy, program, or intervention's design, implementation, or results - be that over an extended period of time or in short cycles that support implementation, operation, adjustment, adaptation, and learning. More simply stated, evaluation is a tool to learn about the value of things we do and, in human services, programs we run. Or even more succinctly, it is a means for testing reality.[1] In this paper, we focus on evaluation as used in relation to human services generally, and youth services in particular.

But first, we need to spend a bit of time considering performance management - because ultimately evaluation has meaning only in the context of (a) trying to understand how well organizations and programs are functioning, (b) whether they are achieving the results they promise, and (c) providing essential support to them by asking essential questions and producing key data.

[1] Patton, M. Q. (2008). Utilization-Focused Evaluation (4th edition). Sage.

Even today there is little understanding that performance management and evaluation are complementary – and especially that evaluation might best be viewed as an essential tool of program performance management.

Performance Management Versus Evaluation – and the Kinds of Data They Use

Evaluation and performance management have had something of an uneasy relationship.[2] For the purposes of CTOP's social investing we rely on the following definitions of performance, performance management, and performance data as applied to programs and their operations:[3]

- **Performance** consists of the degree to which an organization achieves its objectives – and more specifically, how it creates social value. For programs focused on outputs alone, value consists of the quality of the activities and products delivered. For outcome-focused programs, value consists of measurable change for the better in some socially relevant domain.

- **Performance management** consists of "…the set of self-correcting processes, grounded in real-time data measuring, monitoring, and analysis, that an organization uses to learn from its work and to make tactical (front-line, quotidian) and strategic adjustments to achieve its goals and objectives." [4]

- **Performance data** consist of those metrics an organization measures, monitors, and uses in the course of its daily work to keep program quality and effectiveness as high as possible – and if necessary to improve it.

In this connection, it is perhaps worth recognizing that while some evaluators like Ray Rist[5] and Harry Hatry[6] have embraced performance management and performance data as complementary to evaluation and its findings, others are skeptics[7] and still others are openly hostile, dismissing performance data as simplistic and crude.[8] In the nonprofit sector, where there was a prolonged resistance to evaluation and great reluctance among foundations to share findings from evaluations of their programs, **even today there is little understanding that performance management and evaluation are complementary** – and especially that evaluation might best be viewed as an essential tool of program performance management.

---

[2] Bohni, S. N. & Hunter, D. E. K. (eds.) (2013). Performance Management and Evaluation. New Directions for Evaluation. 137.
[3] In general, we use the approach provided by The Performance Imperative: A Framework for Social Sector Excellence (2018) developed by the Leap of Reason Community; see a digital version at: https://www.leapambassadors.org/continuous-improvement/
[4] Hunter, D. E. K. & Bohni, S. N. (2013a). Performance Management and Evaluation: Exploring Complementarities. Ch. 1 in Bohni, S. N. & Hunter, D. E. K. (2013), pp. 7-17.
[5] Rist, R. C. (2006). The "E" in monitoring and evaluation – Using evaluative knowledge to support s results-based management system. In Rist, R. C. & Stane, N. (eds.). From Studies to Streams. Managing Evaluative Systems. Transaction Publishers.
[6] Hatry, H. P. (2013). Sorting the Relationships Among Performance Measurement, Program Evaluation, and Performance Management. Ch. 2 in Bohni, S. N. & Hunter, D. E. K. (2013a). Op. cit. pp. 19-32.
[7] Blalock, A. B. (1999). Evaluation research and the performance management movement: From estrangement to useful integration? Evaluation 5(2). pp. 117-149.
[8] Greene, J. (1999). The inequality of performance measurements. Evaluation, 5(2). pp. 160-172.

What are examples of this complementarity? Three kinds of evaluative work stand out:

First and most fundamentally, evaluators are highly skilled at helping organizations develop "theories of change" – that is, blueprints for building the programs they will use to achieve the results they desire along with the organizational systems and processes for implementing and managing them. As we will see later, theories of change really are at the heart of the matter for managing the performance of social services as well as how they can and should be evaluated. And very few service providers have adequate ones.

> Evaluators provide essential external perspectives to help social service organizations question their assumptions, review their program designs and delivery methods, and pressure test whether their outcomes really are attributable to the efforts of program staff.

Consider the case of the Youth Villages Transitional Housing Program[9], which was developed to help young people with histories of foster care or criminal justice involvement to make successful transitions to adulthood; it provides housing stability and intensive, clinically focused case management, support, and counseling. What outcomes did the young people achieve? They were mixed:[10] "The program boosted earnings, increased housing stability and economic well-being, and improved mental health." However, contrary to Youth Villages' expectations, "it did not increase educational attainment, improve social support, or reduce criminal behaviors" among its former residents. As Gordon Berlin, the former president of MDRC commented recently, this is hardly surprising since the Transitional Housing Program's theory of change conspicuously lacked the kinds of elements that would be necessary to drive educational attainment and reduction in criminal behavior. The bottom line being that neither performance management nor evaluation can be undertaken meaningfully without a strong theory of change.[11]

Secondly, evaluators have a full range of methods with which they can help social service providers measure and monitor the effects of what they are doing – that is, gauge their outcomes. Again, this has little meaning if it is not tied to the desire to manage and if necessary improve program performance.

Thirdly, evaluators provide essential external perspectives in terms of which they can help social service organizations question their assumptions, review their program designs and delivery methods, and challenge the organization to undertake comparative analyses to pressure test whether the outcomes they are tracking really are attributable to the efforts of program staff or whether there might be alternative ways of explaining them. (At a most basic level, external evaluators can help organizations learn about how valid the metrics are that they use and how reliably they are applying them.)

Refreshingly, there is increasing recognition among foundations[12] that both evaluation and performance management are essential parts of what foundations and their grantees need in order to be successful – and to be credible to policy makers as well as the general public. However, that recognition still is not the same as grasping and valuing the complementarity of evaluation and performance management.

Well then – why bother to measure performance in the first place? The noted "performance guru" Bob Behn[13] lists the following reasons (which I have edited down a bit):

- To assess how well the organization is functioning – or, to what degree it is achieving its goals and objectives;[14]
- To manage – that is, to make sure staff are performing the work the way it is supposed to be done;

[9] Valentine, E. J., Skemer, M. & Courtney, M. E. (2015). Becoming Adults: One-Year Impact Findings from the Youth Villages Transitional Living Evaluation. MDRC.
[10] Ibid. p. 99. It is worth mentioning that subsequent to this evaluation the program was renamed to YVLifeSet.
[11] Berlin, G. (2021). Personal communication.
[12] Boris, E. T. & Winkler, M. K. (2013). The emergence of Performance Management as a Complement to Evaluation Among U.S. Foundations. Ch. 6 in Bohni, S. N. & Hunter, D. E. K. (2013a). Op. cit. pp. 69-80.
[13] Behn, R. D. (2004). Multiple Performance Measures. Bob Behn's Management Report. 1(12). Harvard University.
[14] Gordon Berlin (Op. cit., personal communication) emphasizes that this element is essentially empty without evaluation-derived data, a point that I regard as of critical importance; so we're back to the issue of complementarity again.

- To budget, including the need to develop an understanding of the true cost of services provided;
- To motivate, that is, to create performance targets toward which staff should work and receive appropriate acknowledgement when they reach them;
- To advertise the work of the organization to funders, the public, and policy makers, and do so using measures of value that make sense to them and satisfy their requirements;
- To learn – from unexpected successes, from unexpected failures, and from unexpected data;
- To improve, which an organization can't do unless it knows how it has not fully measured up, what went wrong, and what can be adjusted to improve performance going forward.

Knowing the reasons for measuring performance is one thing. Knowing what the measures should be and what to do with them is quite another.

Let's begin by reminding ourselves what performance management is, namely, the set of self-correcting processes, grounded in real-time data measuring, monitoring, and analysis, that an organization uses to learn from its work and to make tactical and strategic adjustments to achieve its goals and objectives. To do so, an organization and its program(s) need to have the following:[16]

- A clear mission with an operational definition of success,
- Measurable goals and objectives in terms of which success will be evaluated,
- A set of Key Performance Indicators (KPIs) that specify areas of performance that are crucial to achieving success,
- A set of measures for assessing performance for each KPI,
- A timetable for measuring and monitoring tactical KPIs in ways that support performance management using real time data (and for tracking strategic KPIs at appropriate intervals),
- Regularly scheduled meetings in which data are analyzed and turned into actionable information by identifying areas of under-performance and using the four questions laid out in Box 1 to develop timely plans for adjusting and improving performance,
- Timelines for implementing the plans and reviewing progress with responsibilities and accountability clearly specified,
- Means for giving public recognition to those who drive performance improvement and celebrating successes; and
- An organizational culture that values data and what can be learned from data.

Box 1 The Four Questions that drive performance management[15]

Question 1:
What do we need to do better?

Question 2:
What do we need to do more of?

Question 3:
What new approach should we try?

Question 4:
What should we stop doing that isn't working (or actually is causing harm)?

In this regard, unlike in purely evaluative work where metrics inherently must be as precise and valid as possible, the same isn't true for the metrics of performance management. As Behn reminds us:[17]

> Don't go looking for the perfect performance measure. Don't spend countless meetings debating whose measureis without defects. Don't hire expensive

[15] For a wealth of advice regarding the use of data to drive performance, consult Behn, R. D. (2014). The PerformanceStat Potential: A Leadership Strategy for Producing Results. Brookings Institution Press.
[16] Behn, R. D. (2006). Performance Leadership: 11 Better Practices That Can Ratchet Up Performance. Second edition. IBM Center for the Business of Government.
[17] Behn, B. (2009). No Perfect Performance Measure. Bob Behn's Management Report. 6(6). Harvard University.

consultants to create the penultimate measure.

Instead, start with a good measure (or two). Not great, not perfect, just good. But from the beginning, try to identify its inadequacies. Recognize what problems the measure might create; then, as you implement your performance strategy, be alert for the emergence of such flaws.

Building on Behn's advice, it is worth making the point explicitly **that performance metrics collected and monitored internally by programs rest on two assumptions**, neither of which may in fact prove to be true once a program is evaluated externally, but both of which are essential for program management.
The first is that the theory of change within which the program was designed and in terms of which it is implemented is justifiable, by which is meant that it will lead to intended results if implemented correctly.
The second is that the metrics themselves are valid (i.e., accurately represent what in general parlance is called the "real world") – one simply has to rely on their "common sense" or "face" validity. As we will see below, an implementation evaluation by an external evaluator using validated metrics will test both these assumptions.

Having located evaluation within a performance management framework, for the rest of this article we discuss evaluation itself – but do so with its applicability to performance management among social service providers always in mind.[18]

Box 2 Mantras for performance management and evaluation

Optimal performance management is not a matter of top-down command and control; rather, it consists of guided and supported, mission-driven self-management at all levels of an organization.

and

Program evaluation that is not linked to performance management is a waste of time, money, and opportunity.

[18] Hunter, D. E. K. & Bohni, S. N. (2013a). Op. cit.

Understanding Evaluation

Probably the best single introduction to this topic is Evaluation: A Systematic Approach (8th edition) by Peter Rossi and his colleagues.[19] They say that program evaluations must be tailored to local circumstances and the needs of key constituent groups, and discuss five domains where evaluators can be of great value to programs and the organizations that operate them:

1. Performing a program needs assessment,
2. Helping with program design,
3. Assessing program implementation and service delivery,
4. Conducting program outcomes and impact studies and
5. Assessing program efficiency.

In this country, it was in the fields of health and education during the 1960s that evaluation emerged as a professionalized line of work. One of its most influential practitioners early on, along with Peter Rossi,[20] was Michael Scriven[21,22,23,24] and in no small part due to his influence for some two decades evaluation was mostly used for the purpose of establishing program effectiveness and creating means to hold programs accountable for their results. Basically, evaluation was used to answer relatively straightforward questions: Did the program produce the effects it promised - did it work? Was the cost of the Initiative justifiable in terms of its outcomes? These are high stakes questions, and evaluations that answered them often had profound consequences: based on their findings, government programs were expanded or discontinued, funds were redirected, foundations created new revenue streams and at times were quick to eliminate them - often because policy makers and funders in general wanted quick answers, didn't fully understand the theory behind and the mechanics of evaluations, over-interpreted evaluation findings, and made decisions on very preliminary data.[25]

When evaluation first emerged as a professionalized line of work in the United States in the 1960s, it was used to answer relatively straightforward but high-stakes questions, and evaluations that answered them often had profound (and not necessarily good) consequences.

[19] Rossi, P. H., Lipsey, M. W. & Henry G. T. (2018). Evaluation: A Systematic Approach (8th edition). Sage.
[20] Rossi, P. H., Freeman, H. & Rosenbaum, S. (1979). Evaluation: A Systematic Approach. Sage. This was the first widely used textbook on evaluation.
[21] Tyler, R. W., Gagne R. M., & Scriven, M. (eds.) (1967). Perspectives of Curriculum Evaluation. Vol 1. American Educational Research Association Monograph Series on Curriculum Evaluation. Rand McNally.
[22] Scriven, M. (1974). Evaluation: A study guide for educational administrators. Nova University.
[23] Scriven, M. (1982). Logic of Evaluation. EdgePress
[24] Scriven, M. (1987). Theory and Practice of Evaluation. EdgePress.
[25] Unfortunately, these occasions stimulated what I think of as a Luddite reaction among many nonprofit and foundation leaders who blamed all attempts to use evaluation for the stupid ways in which it was misused by its consumers, themselves included.

An example of the latter was the decision by the George W. Bush administration in 2003 to cut about $400 million from the $1 billion federally-funded 21st Century Community Learning Center Program[26] in response to the 2002 release by Mathematica of its preliminary findings after one year of a planned three-year impact evaluation. Why? Because the data showed no academic gains by the participating children. Admittedly, following considerable public blowback supported by scorching commentaries from evaluators (some of whom harshly criticized the study's methods while others criticized the politicians for rushing to judgement[27]), the Bush administration restored the cuts – but not before the emergence of a groundswell of anti-evaluation sentiment in the nonprofit social services sector that, it should be noted, survives with many adherents to this day.

Alternatively, because for a long time many (but of course not all) evaluations tended to focus on impacts rather than on program designs and delivery methods that produced or failed to produce them (evaluators frequently relegated these matters to a "black box" and simply studied what went in and what came out – but not what was inside), practitioners, funders, and even policy makers often consigned evaluation reports unread to dusty shelves or rarely-opened file cabinets. Impact evaluations simply were not designed to answer questions about how things worked or failed to, and how they might be improved. And in truth, some outcome and impact evaluation reports are delivered to the evaluated programs too late to be of practical use and are too academic to be helpful, too dense to be interesting or even accessible to the lay person, and too expensive to boot.

Furthermore, grantees often find foundations' evaluation requirements exceedingly burdensome, of no practical value, and a drain on resources.[28]

So, by the 1970s it was becoming apparent that using evaluations only for high-stakes assessments to prove (or disprove) program effectiveness and promote accountability was missing the boat in some very important ways. Mostly, since evaluations were inherently backward-looking (did it work?) and therefore of little use to people working on the ground, individuals and groups who wanted to learn in the present how to improve the quality and effectiveness of their efforts looked for and developed alternative evaluative methods. Books by Michael Quinn Patton[29] and Chen Huey-Tsyh[30] were linchpins that moved evaluation away from a preoccupation with accountability and toward a concern with practical applications that supported learning and utilization.[31] Related to this, some evaluators began to emphasize the importance of qualitative data (alongside the usual quantitative data) as inherently necessary to understand programs, how they work, and the results they achieve.[32]

Interestingly, this shift in direction amounted to a trip back to the future.

The first evaluation on record

Some 2,800 years ago, in 605 BCE to be precise, King Nebuchadnezzar of Babylon conquered the people of Israel and, as the Bible tells us, had thousands of them brought back to his capital as servants.[33] Of these, he commanded that a group of Israelite aristocrats be instructed in the Babylonian language, in its arts, and in its store of knowledge so that they could serve him as mid-level bureaucrats. And to make sure they would thrive, he insisted that they be given the very same food he ate.

However, among this select group was a man by the name of Daniel who, along with three friends, could not bring themselves to eat the king's food because it wasn't Kosher. They threatened a hunger strike and

[26] This program supports the creation of community learning centers that provide academic enrichment opportunities during non-school hours for children, particularly students who attend high-poverty and low-performing schools. The program helps students meet state and local student standards in core academic subjects, such as reading and math; offers students a broad array of enrichment activities that can complement their regular academic programs; and offers literacy and other educational services to the families of participating children.
It continues to be a major source of funding for after-school (or extended day) programming.
[27] See for example: Bissell, J. S., Cross, C. T., Mapp, K., Reisner, E., Vandell, C. W., & Weissbourd, R. (2003). Statement released by the Scientific Advisory Board for the 21st Century Community Learning Center Evaluation, May 10th.
[28] Patrizi, P. & McMullen, B. (1999). Realizing the potential of program evaluation. Foundation News and Commentary. 40(3) pp. 30-35.
[29] Patton, M. Q, (1979). Utilization-Focused Evaluation (1st edition). Sage.
[30] Chen, H.-T. (1990). Theory-Driven Evaluations. Sage.
[31] Two of America's most prominent evaluation shops emerged around this time: The Manpower Demonstration Research Corporation (now officially renamed MDRC) was founded in 1974 and Public/Private Ventures (P/PV) in 1978, both of which were partly funded by multiple U.S. government agencies as well as the Ford Foundation to inform, design, implement, and evaluate strategies to help improve the lives of people living in poverty.
[32] Patton, M. Q. (1980). Qualitative Evaluation Methods. Sage.
[33] A fuller discussion of these events is in Hunter, D. E. K. (2006). Daniel and the Rhinoceros. Evaluation and Program Planning, 29, 180-185.

Program outcomes are the expected, measurable changes undergone or achieved by program participants. Usually these consist of changes in attitudes, knowledge, skills, behavior, status (e.g., graduating from school or obtaining employment), and social or personal condition.

when their supervisor let them know his own life could be on the line if they acted on the threat, Daniel proposed a dietary evaluation in which he and his colleagues would be allowed to eat a porridge of legumes and thus avoid eating what to them was "unclean" meat. Then he said, as reported in the King James version of the Bible, "... let our countenances be looked upon before thee, and the countenances of the children [of Israel] that eat a portion of the king's meat: and as thou seest, deal with thy servants." So he made it a high-stakes evaluation and, when Daniel and his friends thrived, they were allowed to continue with their diet. Daniel went on to outperform all the other captive Israelites including those who ate the king's meat. In fact, during the reign of Nebuchadnezzar's son and successor Belshazzar Daniel, reading the fiery handwriting on the wall foretold the subsequent fall of Babylon to the Persians.

So what do we think of this evaluation?

Well, from a scientific perspective it's pretty weak...actually it's terrible because four subjects can't possibly provide a statistically significant result. Further, since they selected themselves, we have here a clear case of what evaluators call selection bias. Add to this that the intervention was so short-termed that it is highly unlikely to have affected the findings one way or the other; and that the outcome measures were impressionistic, utterly subjective, and pretty unlikely to have been valid.

But it is worth noting that the evaluation also had some merits. The intervention it assessed had a clearly formulated logic that all stakeholders accepted; the evaluation's design was created in an inclusive manner that involved key stakeholders; and it was definitely intended to be used – it would inform subsequent policy decisions. So here we have an evaluation that, while at first glance is problematic, is not entirely so. It was designed to be useful, was inexpensive, built local evaluation capacity, supported high-stakes decision-making, accomplished what it was designed to do, and was useful to all involved. It is worth keeping these points in mind as we move on to think about how evaluation can be useful to front-line service providers, to the funders who support them, and to the policy makers whose decisions affect them so profoundly.

But before we dive into the broader matters of evaluation, it's essential that we first make clear what is meant when we use the term "program outcomes".

Program Outcomes

Program outcomes are changes shown by program participants that, it is hoped and expected, are the results of the program's activities.[34] This requires us to make a key distinction between program outputs and program outcomes.

Program outputs typically are the set of activities in which staff (and/or volunteers) engage, the number of people they serve, the number and percent of people served who actually complete the program, and the kinds of materials they produce and/or disseminate (e.g., materials for use by workshop participants or to distribute to the public).

On the other hand, **program outcomes** are the expected, measurable changes undergone or achieved by program participants. Usually these consist of changes in attitudes, knowledge, skills, behavior, status (e.g., graduating from school or obtaining employment), and social or personal condition (e.g., becoming a parent or shifting from antisocial to prosocial friendships, from being homeless to being domiciled, from not working to full employment, from being sick to being well).

It is useful to distinguish among three kinds of outcomes:

- **Long-term outcomes** are what program participants achieve after they have left a program for a period of time; **these are the measures of a program's social value.**

[34] Hunter, E. E. K. (2013b). Working Hard - and Working WELL. Hunter Consulting, LLC. pp. 74-76.

• **Intermediate outcomes** are observable changes in participants that are monitored periodically to ascertain whether they are progressing in a timely way (as called for in the program's design) and whether all subgroups of participants are benefitting equivalently. The final or ultimate intermediate outcome(s) should consist of the indicators used by a program to determine that, as they leave it, graduates are ready, willing and able to achieve the long-term outcomes that lie ahead of them.

• **Short-term outcomes** are the small gains program participants make (while still in the program) in direct response to program activities – they can be thought of as rungs on a ladder leading to intermediate outcomes.[35]

Program outcomes, to be relevant for the delivery of social services, must be:

- Socially meaningful,
- Measurable and monitored,
- Sustained,
- Logically linkable to the program's activities, and
- Comprise that for which stakeholders (participants, staff, management, funders, local residents, etc.) hold the program accountable.[36]

The following dictum can't be emphasized enough:

Outcomes can't be bought. Money buys outputs. Smart, intentional, and focused management of outputs produces outcomes.

And now, as promised, to the heart of the matter.

Outcomes can't be bought.
Money buys outputs.

[35] Not all evaluators use these three terms in this manner. Some locate outcomes after program participants have graduated, in which case short-term outcomes would be manifested immediately or shortly after exit, and intermediate and long-term outcomes at appropriate intervals thereafter. In this approach, even short-term outcomes can be large – e.g., very significant gains in income. I appreciate Gordon Berlin (2021; Op. cit.) drawing my attention to this matter.
[36] Hunter, D. E. K. (2013b). Op. cit. p. 124.

Theory of Change: The Heart of the Matter

Before we proceed, it is necessary to recognize that unless a program and the organization providing it have a theory of change it will be close to impossible to deliver the program reliably and effectively, and then to evaluate the program's performance meaningfully.[37] What, then, is a theory of change[38]? Simply put, a "... theory of change is best thought of as an organization's blueprint for success. It is the guide whereby the organization plans, structures, and engages in its daily activities to achieve its strategic goals and objectives – and in particular, its intended results. It also provides the framework within which an organization can examine what works and what does not work within its... programming, and manage performance for continuous improvement."[39]

A well thought-through theory of change will answer the following questions among others:[40]

- Whom is the program intended to serve and benefit (target population)?
- What kinds of evidence and other considerations (such as stakeholder perspectives) informed the program design?
- How does the program design address the needs and desires of its target population?
- What is the program model? What are its elements, its activities? Where, how frequently, how intensively, will the program engage participants in these activities (often called "dosage")? Over how long a time period?
- What does the program have to monitor to ensure that it is operating at a high level of quality?
- What are the immediate or short-term outcomes that are measured and tracked to learn whether clients are benefitting incrementally in a timely way?
- What are the intermediate outcomes that are monitored to determine whether program participants are progressing toward long-term outcomes in a timely way?
- How will the long-term outcomes be tracked and measured? And what is their social value?
- What percentage of program "graduates" are expected to achieve and sustain the program's long-term outcomes?
- How many people can the program serve at a given time so that the full range of activities can be delivered at the right dosages for the length of the program?
- How is the enrollment of participants managed to ensure that enrollees are members of the target population?
- What are the external constraints that might interfere with clients benefitting from the program as intended?

---

[37] Chen, H-T. (1990). Op. cit.
[38] Dhillon, L. & Vaca, S. (2018). Refining theories of Change. Journal of Multidisciplinary Evaluation 14(3). pp. 64-87. Admittedly a bit uneven in its thinking and presentation, especially where the article seems to conflate long-term outcomes with impacts, it nevertheless brings up interesting thoughts worth considering with regard to theories of change and their use.
[39] Hunter, D. E. K. (2006) Op. cit. p. 183.
[40] Hunter, D. E. K. (2013b). Op. cit.. pp. 42-43. Admittedly my use of the term is broader than those of many evaluators, for whom a theory of change is pretty much equivalent to a program logic model. However, since programs are delivered by organizations that are a causal context with regard to their design, implementation, and management, I think it is essential to include such broader considerations as part of the organization's theory of change; in this view program logic models are but one element of the whole.

And what systematic steps is the program or the larger organization within which it works taking to meliorate these constraints?

How do we know if a theory of change is good? Henry Mayne, an important thought leader in the area of evaluation, offers the following quality indicators:[41]

- **It should be plausible.** Does common sense suggest that the activities, if implemented, will lead to desired results?
- **It should be agreed.** Is there reasonable agreement with the theory of change as postulated?
- **It should be embedded.** Is the theory of change embedded in a broader social and economic context, where other factors and risks likely to influence the desired results are identified?
- **It should be testable.** Is the theory of change specific enough to measure its assumptions in credible and useful ways?

To this list it is useful to add that a theory of change should be:[42]

- Doable within resource constraints, and
- Operational, that is, it provides a useful framework for managing organizational performance reliably, sustainably, and at high levels of quality and effectiveness – that all staff agree with and use to monitor and manage their work (and, in the highest performing organizations, to hold each other accountable for high performance).

From all this it should be clear why any program needs a theory of change. And absent one, evaluating it would be about as productive as measuring the combined length of the pasta in a plateful of spaghetti.

> Unless a program and the organization providing it have a theory of change it will be close to impossible to deliver the program reliably and effectively, and then to evaluate the program's performance meaningfully. Therefore, absent a theory of change, evaluating a program would be about as productive as measuring the combined length of the pasta in a plateful of spaghetti.

[41] Mayne, J. (2008). Addressing Cause and Effect in Simple and Complex Settings through Contribution Analysis. Discussion draft in: Schwartz, R., Forss, K., & Marra, M. (eds.). Evaluating the Complex. At the time, forthcoming.
[42] Hunter, D. E. K. (2013b). Op. cit. p. 50.

Four Major Functions of Evaluation

This paper is not meant to be an academic treatise on evaluation, nor to be encyclopedic. Its purpose is to examine some ways in which evaluation can and should be used by and for funders of social services and their grantees - and indeed the ways in which we think about and use evaluation as part of our work in the Connecticut Opportunity Project.[43]

The four major functions of evaluation discussed for this purpose are:

1. Assessing program outcomes;
2. Establishing program impact;
3. Assessing program fidelity of implementation; and
4. Informing program development, implementation, improvement, management, and what can be learned from it.

Function 1: Assessing Program Outcomes

Using the definition of outcomes presented above, it should be clear that there is considerable value in evaluating a program's outcomes. An outcome evaluation, therefore, "...investigates whether changes occur for participants in a program and if these changes are associated with a program or an activity. Such evaluations examine whether, to what extent, and in what direction outcomes change for those in the program." [44]

In studying program outcomes it is important to decide how they will be calculated: Are they to consist of the extent of change (e.g., improved reading scores), as an interim metric, or a final condition or status (e.g., graduating from high school, becoming domiciled), or both. So one must be clear about the program's objectives and also on the metrics that would be meaningful to its providers and other stakeholders (especially program participants) for assessing the outcomes that the program is intended to achieve as specified in the theory of change.[45]

[43] CTOP is an initiative of Dalio Education, which in turn is a division of the Dalio Foundation.
[44] Allen, T. and Jacinta Bronte-Tinkew, J. (2008). Outcome Evaluation: A Guide for Out-of School Time Practitioners. Series on Practical Evaluation Methods. Child Trends.
[45] Ibid. p. 8.

Function 2: Establishing Program Impact

Not all human services programs are intended to produce outcomes. Many are conceived to provide such things as enrichment activities or recreational opportunities; to distribute essentials such as free food to people who are homeless; or to disseminate important information for the general public (e.g., information on how to cope with the Covid-19 pandemic). These kinds of programs often are called **output programs, and their tasks are (a) to ensure that their outputs are of the highest possible quality and (b) that they attract the largest possible number of consumers or users.**

On the other hand, many social programs are designed, funded, and delivered for the purpose of creating socially important changes - in individuals, families, groups, neighborhoods, in life outcomes, etc. They are meant to accomplish things like improving adult literacy, improving children's school performance, preparing people for success in post-secondary education and/or the workforce, reducing levels of violence in particular neighborhoods, improving the benefits to patients of health services, reducing the prevalence of driving while under the influence of drugs or alcohol, reducing recidivism among prisoners returning to community-based living, reducing the rate of teenaged pregnancy and parenthood, promoting civic engagement, and on and on.

For such programs, it is essential to know whether they really are delivering the results they promise compared to what we might call "life as usual". First, because if they aren't, they are wasting a lot of time and money. Second, they potentially are attracting program participants away from programs that do deliver results - which is captured under the concept of "opportunity cost." And third, at least for some participants, the risk is high that in experiencing yet another failure in a program that doesn't work they will become less hopeful, more demoralized, and more alienated.

Because this issue is so important, it seems like a good idea to address it head-on: How can one establish whether or not a program is really making a difference, and is having any impact(s) on its participants? And the answer is that this requires an impact evaluation. And an impact evaluation requires the use of a counterfactual.

To repeat: **Impacts are outcomes that have been proven, using methods that meet scientific standards**[46] **, to be the results of a program compared to what would have happened to participants otherwise.**[47] How can one accomplish this? By comparing the outcomes of program participants to people who are very much like them in all ways that are significant but who did not participate in the program, and measuring these people in terms of the same outcomes for which program participants are assessed. Thus the people to whom the outcomes of program participants are compared constitute the counterfactual component necessary for any impact evaluation.

To be sure, any effort at determining causes involves paying attention to some factors (such as a program, its participants, its elements) and ignoring many others (contextual factors ranging from prevailing cultural norms to sociopolitical realities, economic realities such as the local availability of jobs, family composition, local health patterns, and even the climate). In other words, any causal statement inevitably involves a profound simplification of reality. Nevertheless, one has no choice but to try, within the limits of our understanding, to evaluate whether programs are delivering the results

[46] The generally accepted standard is that the evaluation's results have a 95% or better probability of being valid.
[47] Often the term "impacts" is used in common language to refer to long-term outcomes, even though they have not been established as such through an impact evaluation. Impacts, as used herein, is what may be thought of as a term of art for professionals.

they promise. **Use of a counterfactual, in which people are as similar as possible in all regards to the program participants (who are often called the "experimental group"), is the main method for doing so in that, to a large degree, contextual issues (and other variables that might not be obvious) are essentially the same or very similar for both groups.** And while this may be imperfect, it is the best we can do.[48]

Judith M. Gueron, then-president of MDRC, wrote about a relevant case:[49] the evaluation of three programs intended to help mothers on welfare get and keep employment in the workforce. For the evaluation, eligible mothers were randomly assigned to one of the three programs or to the matched control groups where they didn't receive any services. Looking at the results, the program in Grand Rapids showed an employment rate for program completers of 67 percent; the program in Atlanta had a rate of 57 percent; and the program in Riverside, California had a rate of 46 percent. So it would appear at first glance that the Grand Rapids program was the most effective, and the Riverside program the least. However, when looking at the programs' impacts by comparing their success rates to the success rates of their respective counterfactual control groups, the results were tuned upside down: it turned out that the Riverside program had the highest impact in that it improved employment by 8 percent; the Grand Rapids program was in the middle at 6 percent; and the Atlanta site had the lowest impact at 5 percent.

At this point it is worth reminding ourselves that there are many examples of programs that, when studied comparatively, show no impacts. One notable example is D.A.R.E., the much-loved and highly funded police-run drug abuse prevention program. Multiple impact studies have proven time and again that it simply doesn't work.[50] Furthermore, impact studies of some programs have actually revealed that they do harm. One illustration is Scared Straight, which is intended to prevent juvenile criminality by bringing at-risk young people to visit prisons and show them the horrors of prison life. "Far from reducing crime, research shows that participants in Scared Straight programs are about 7 percent more likely to commit crimes afterward than those who don't participate. This finding is not even new – repeated studies carried out since the 1960s show that Scared Straight programs have no positive effect." [51]

There is another purpose served by the use of a counterfactual, which is to manage sources of bias that can distort the results of an evaluation. Of the many possible sources of bias, three are especially important: maturational bias, selection bias, and history bias.

Maturational bias is seen where outcomes that are attributed to a program really would have come about for participants even without it, simply as a fact of human development. Thus babies in a day care center will, over the course of a year, get longer and heavier. If the program selects these metrics for their outcomes, they are biasing their so-called results massively toward success.

Selection bias comes about because the enrollment process (wittingly or unwittingly) recruits many participants who already have achieved the intended changes when they were enrolled – or who have a better than average likelihood of achieving them. A case in point: I was asked to review a program in the U.K. intended to improve high school graduation rates for children who showed key risk factors for dropping out of school prematurely. But when I looked closely at the program, it turned out that front line staff in the schools were deliberately recruiting very successful students onto their caseloads because they knew the program as designed couldn't possibly work for at-risk students, and at the same time program management had imposed on them a demand for an 80 percent high school graduation

When evaluating programs to learn about their impact, the use of a counterfactual is the only way to manage sources of bias that can distort the results. Three sources of bias are especially important: maturational bias, selection bias, and history bias.

[48] Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin.
[49] Gueron, J. M. (2005). On the Frontlines Throwing Good Money After Bad: A common error misleads foundations and policy makers. Stanford Social Innovation Review. Fall. pp 69-71.
[50] West, S. L. & O'Neal, K. K. (2004). Project D.A.R.E. Outcome Effectiveness Revisited. American Journal of Public Health, 94(6). pp. 1027-1029.
[51] Kohli, J. (2012). Why Scared Straight Programs are a Waste of Taxpayer Dollars. Doing What Doesn't Work. Center for American Progress.

rate for the young people they were serving. This is a supercharged example of what often is called "skimming" or "creaming" – selection bias on steroids.

History bias exists where contextual factors change significantly during the time that a program is being evaluated, and these changes significantly improve or diminish participants' likelihood of achieving intended outcomes. For example, a prison release (reentry) program that shows declining rates for reincarceration among its graduates could simply be benefitting from the well documented drop in crime rates nationally over the past thirty years. Or consider the fact that while crime did start to fall after the Clinton administration funded an increase in police officers nationwide in the 1990s, the crime rates continued to fall even when the number of police started dropping again. So in fact, while advocates of "law and order" policies such as increasing police may point to examples such as these, serious evaluators and policy analysts have refuted their assertions and shown that the drop in crime rates was caused by the confluence of multiple historical variables – and that no single policy or approach or program drove it.[52]

The use of a counterfactual is the only way we know how to deal with these biases when evaluating programs to learn about their impact. The questions that arise, then, are (a) what are the methods for constructing a counterfactual; (b) when should programs undertake evaluations that use them; and (c) what are some of the usual challenges that arise in the process?

What are the main methods for designing a counterfactual to evaluate program impact?

In this section we consider two[53] approaches to establishing a counterfactual to study a program's impact:

- Randomized controlled trials (RCTs)
- Quasi-experimental methods

We also discuss benchmarking, because we see such studies as demonstrating that programs are producing outcomes – but also see them as falling short of proving impact.

Randomized controlled trials (RCTs)

RCTs often are called the "gold standard" of program evaluation, with the suggestion that since they use "experimental methods," any other methodology is less than adequate for establishing program impacts. But in reality, like all evaluation designs, RCTs are a compromise among competing factors that vary from one context to the next, from one program to the next, from one participant group to the next, from one set of stakeholders to the next, and even from one evaluator to the next. I believe that designating RCTs as the "gold standard" is not helpful since it distracts from the consideration of the compromises that have been made in any given case, and the reasons that lie behind them.[54]

What distinguishes an RCT from other evaluation methodologies is the use of a control group – that is, a group of individuals who have been selected in such a manner that they are virtually indistinguishable in their characteristics from a profile of participants in the program being evaluated, and whose outcomes will be compared to those of the individuals in the program (the so-called "experimental group"). The manner

What distinguishes an RCT from other evaluation methodologies is the use of a control group – that is, a group of individuals who have been selected in such a manner that they are virtually indistinguishable in their characteristics from a profile of participants in the program being evaluated.

[52] Behn, R. D. (2014). The PerformanceStat Potential: A Leadership Strategy for Producing Results. Ash Center for Democratic Governance and Innovation; Brookings Institution Press.

[53] Many people would add benchmarking to this list. However, at CTOP we regard benchmarking as a means for demonstrating program effectiveness, rather than proving it to the level where we would credit a program's outcomes with being impacts. When benchmarking is used to look at a program's effectiveness, they remain for us "demonstrated outcomes."

[54] It is worth noting here that RCTs of human service programs have a bias vulnerability built into them by their very nature. Unlike "double blind" studies used, for example, in medical research such as on vaccines where neither the participants nor the evaluators know whether a given individual is in the group receiving the intervention or in the control group receiving placebos against which it is being compared, human service RCT evaluators know full well who is in the program and who is in the control group and consequently are vulnerable to subjective bias in making their measurements or assessments. This vulnerability they have in common with all methods used to investigate program impacts and outcomes. However, RCTs do eliminate the three key sources of bias in evaluations – selection bias, maturational bias, and history bias – better than do any other evaluation methods, and thereby have earned the designation of "gold standard" in many evaluators' views.

of selection is randomization. "The key to a random assignment experimental study is that members of both the experimental group and the control group are, as groups, the same or very similar. It's that simple." [55] And RCTs provide an elegant and minimalist way of making sure that all the variables that could make a difference beyond those used to determine who should be enrolled in a program are equivalent between the experimental group of program participants and the control group – without ever having had to measure them.

Randomization involves identifying a pool of individuals who meet the criteria for enrolling in a given program and then, before any are enrolled, randomly assigning them into either the experimental or control group. In essence, this amounts to the equivalent of using a flip of the coin to determine, for each member of that pool, whether she or he will be enrolled as an active program client or be delegated to the counterfactual group where the standard practice is to provide "service as usual" – that is, letting them go about their lives and make use of other program services in the community as they normally would. But for the duration of the evaluation these individuals are precluded from enrolling in the program under study.

The issue of denying service to some but not others poses ethical considerations. But because the number of slots is limited in most social services programs, those programs in the normal course of events effectively are denying services (but perhaps don't feel it in the same way). They use concepts like first come, first served or limit recruitment or establish elaborate screening and selection criteria to maximize the likelihood that those served will benefit as measured by the program's outcomes. For these reasons, random assignment, in which all of those eligible have an equal chance of getting in, is ethical, even though understandably it makes program administrators and staff uncomfortable. Random assignment is not ethical if the program is an entitlement, in which everyone who wants services or benefits are admitted. It is also helpful to remember that the evaluation is being done because we don't have reliable evidence that the program is having the intended effect. Indeed, positive results can lead to an increase in resources that enables a much larger number of people to be served.[56]

It is undeniable that RCTs have been of great value to practitioners in social services, and to their funders as well.[57] As mentioned above, they have identified programs that simply don't work – for example New Chance, a program for young mothers and their children.[58] And some have shown that while programs are effective for some subgroups of participants, they are not helpful for others – an example of which is the Carrera Pregnancy Prevention Program of New York's Children's Aid Society, which changes girls' sexual activities but not those of boys.[59] On a positive note, some programs that initially showed no impacts were discovered, through RCTs, to have important impacts on former program participants many years later – such as the Perry Preschool Program.[60]

Nevertheless, for many practitioners it is unsettling to let chance drive whether or not a given person can enroll in a program; and there can be strong resistance to the idea that it will be necessary to keep individuals out of a program for an extended period of time even when program slots open up and staff believe they could benefit from the program.[61] As noted by the evaluators Kristin Moore and Allison Metz,

> A program provider who works hard every day and believes in the value of his or her program is naturally going to have reservations about assigning children or youth to a control group that is not in the program. Even when random assignment is deemed the most appropriate evaluation approach, many program managers and practitioners feel it is unfair to deny potentially beneficial services to children or their families. This is completely understandable.[62]

What to do when you can't do an RCT? First not every question requires one. As Gordon Berlin[63] points out, "…if it is about implementation or operational matters

[55] Moore, K. A, & Metz, A. (2008). Random Assignment Evaluation Studies: A Guide for Out-of-School Time Program Practitioners. Child Trends Research to Results Brief. p. 1.
[56] Baron, J. & The Coalition for Evidence-Based Policy have produced a very helpful set of factors to look at when considering how well an RCT evaluation has been conducted. See: Coalition for Evidence-Based Policy (updated2010). Checklist for Reviewing a Randomized Controlled Trial of a Social Program or Project, to Assess Whether It Produced Valid Evidence.
[57] Moore & Metz, Op. cit. p. 4.
[58] Quint, J. C., Bos, J. M., & Polit, D. F. (1997). New Chance: Final report on a comprehensive program for young mothers in poverty and their children. New York: MDRC.
[59] Philliber, S., Kaye, J., & Herrling, S. (2001). The national evaluation of the Children's Aid Society Carrera-Model.
Program to prevent teen pregnancy. New York: Philliber Research Associates.
[60] Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). Significant benefits: The High/Scope Perry preschool study through age 27. Monograph of the High/Scope Educational Research Foundation, 10. High/Scope Press.
[61] But of course such beliefs, although held passionately and tenaciously, in fact may not be grounded in reality; strong organizational and program leaders will make this point to staff and argue that it is precisely to determine whether such beliefs are justified that an impact evaluation is necessary.
[62] Moore & Metz, Op. cit. p. 4.
[63] Berlin, G. (2021). Personal communication.

rather than program content, one isn't appropriate – in fact it won't answer the question. When you want to answer the "what difference" question, an RCT often is the best way to do so and can be used in many more instances than commonly thought. But when it can't you do the best you can and you caveat it as needed, which is usually a fair amount. The best protection is multiple quasi-experimental tests over time – if they all point in the same direction, your confidence grows."
Such considerations have led practitioners and evaluators to look for alternative methodologies to create counterfactuals for studying program impacts. Often they involve what are called "quasi-experimental" designs.

Quasi-experimental methods

As with RCTs, the intent is to identify a counterfactual comparison group whose characteristics are as close as possible to the group who is enrolled in the program so that differences in outcomes between the program group and comparison group will not be the result of selection bias. This is done by picking key demographic (age, gender, ethnicity, etc.) and risk factors (living in or aging out of foster care, involved with the juvenile or criminal justice system, dropping out of school or failing core courses, etc.) that characterize program participants – and selecting individuals for the comparison group that are matched point by point with program participants in terms of these factors. The larger the number of matched points, the higher the confidence we can have in the study. "In other words, matching seeks to identify subsamples of program and comparison group members that are 'balanced' with respect to observed covariates: the observed covariates are essentially the same…" [64] The challenge, however, is how to control for those covariates that can't be observed – for example, motivation – that can't be eliminated in quasi-experimental methods the way they are in RCTs.
Commonly used quasi-experimental methods are "before-after" comparisons, interrupted time series comparisons, and regression discontinuity comparisons, all of which use earlier assessments of program participants as the counterfactual for later assessments. Some evaluations construct synthetic comparison groups from longitudinal studies that follow cohorts of similar people over time, for example, the National Longitudinal Survey of Youth. Others compare near similar groups using an eligibility cut off point like age or income, and compare those just barely eligible for the program to those just on the other side of that cutoff point.[65]

The use of a quasi-experimental matched comparison group entails pretty much the same costs as an RCT, since in both cases individuals in the counterfactual group have to be tracked and assessed periodically as called for by the evaluation design. But in fact, in many cases it is quite possible to do low-cost, rigorous RCT program evaluations as well as quasi-experimental evaluations using public data sets.[66] So the reasons for adopting a quasi-experimental approach generally have to do with ethical or practical concerns, not matters of cost, and these have to be considered by the program operator and the evaluator in arriving at a mutually acceptable evaluation design.

Benchmarking: demonstrating program effectiveness

For many reasons, including practicality and cost considerations, it may not be feasible for a program to undertake a scientifically rigorous impact evaluation. In such cases, it would nevertheless be useful and in some cases even imperative to compare program results to a counterfactual, especially if the program is growing. One means for doing so is to use benchmarking, a tool used in many businesses, in the public sector, and by environmentalists as well as by nonprofit program providers to monitor and improve performance.[67]

Like RCTs and quasi-experimental evaluation designs, benchmarking studies seek to demonstrate the effectiveness of programs by comparing their outcomes to a meaningful reference group or program. In human services, this generally involves the use of public data sets or data from other programs doing similar work with similar target populations that show how people resembling the program participants fare with regard to such things as educational attainment, employment, earned income, criminal involvement, age of pregnancy, or civic engagement (often measured via voting behavior). Where there are statistically significant differences in outcomes between the two, this is taken to be a measure of program effects.

[64] Stuart, E. A. & Rubin, Donald B. Ch. 11 in Osborne, J. (ed.). (2008). Best Practices in Quantitative Methods. Sage.
[65] Reichardt, C. S. Quasi experimental design. In Mathison, S. (ed.). (2005). Encyclopedia of Evaluation. Sage. pp. 351-355.
[66] Baron J. & The Coalition for Evidence-Based Policy. (2012). Rigorous Program Evaluations on a Budget: How Low-Cost Randomized Controlled Trials Are Possible in Many Areas of Social Policy.
[67] Rolstadas, A. (2013). Benchmarking - Theory and Practice. Springer Science and Business Media.

> Benchmarking can be internal or external. When benchmarking internally, organizations benchmark against their own projects....External benchmarks are generally considered to provide the greater advantage; however, internal benchmarking can be useful where no external benchmarks are available. Internal benchmarks are often the starting point for quantitative process examination. Trends can be identified by examining these data over time, and the impact of performance-improving processes can be assessed....[But] without external benchmarks, an organization and its managers may lack an understanding of what constitutes "good" performance.[68]

For the purposes of CTOP's social investing strategy [69] we regard benchmarking as a means to demonstrate – but not to prove – program effectiveness. The evidence in benchmarking generally is weakened by issues about the comparability of the programs, groups, or populations being compared, as well as concerns regarding the equivalence of measures used to generate the data via which comparisons are made. Nevertheless, benchmarking is a practical and valuable tool for programs wishing to assess and review their effectiveness at periodic intervals.

So the question arises: What kinds of evidence are relevant to assessing the likelihood that a program is effective? That it works as advertised? That it is benefitting participants in the ways it promises? In other words – that is having the expected impact? Box 3 shows the hierarchy of evidence that CTOP uses to judge how confident on can be that a program actually is beneficial, that it has real societal value in that it is measurably improving its participants' lives and prospects.

By the way, neither an implementation evaluation (discussed below) nor an impact evaluation should be undertaken until a program has been assessed for its evaluability – that is, whether it is being delivered and managed in ways that valid and reliable information about its operation and effects can be obtained.[70]

We close this section with a sober note on evaluations and how they may or may not be used by organizations to drive performance management. "The findings from impact evaluations – and indeed any evaluative activities – will not be used well unless and until we tackle and reform organizational culture. [It is essential to take on]...the task of eliminating disincentives and creating incentives for adopting evaluation findings. Without rethinking the incentive and reward systems that guide the behavior of employees, evaluation is likely to remain a marginal activity as opposed to a core driver of decision-making." [71]

Box 3 CTOP's hierarchy of evidence concerning program effectiveness

CTOP uses four levels to designate the degree of confidence one might have regarding a program's ability to produce good results – impacts – for participants as intended. Our intent is that CTOP grantees will, over the course of our investment, reach Level 2 at a minimum.

Level 0	Level 1	Level 2	Level 3
Asserted effectiveness as supported by anecdotal data	Apparent effectiveness as supported by internally collected outcome data	Demonstrated effectiveness as supported by well benchmarked outcome data	Proven effectiveness as supported by one or more RCT or quasi-experimental impact evaluations

[68] Committee for Oversight and Assessment of U.S. Department of Energy Project Management. (2005). Measuring Performance and Benchmarking Project Management at the Department of Energy. Ch. 3:22. The National Academies Press.
[69] Available at: https://www.ctopportunityproject.org/Customer-Content/www/CMS/files/2020-10-11_CTOP_Strategy_Proposal.pdf. This paper is updated from time to time to reflect adjustments made in light of what we are learning through this work.
[70] Smith, M. Evaluability Assessment. In Mathison, S. (ed.). (2005). Encyclopedia of Evaluation. Sage. pp. 136-139.
[71] Bonbright, D. (2012). Use of Impact Evaluation Results. Impact Evaluation Notes. no. 4. InterAction.

When should programs undertake an impact evaluation?

By their very nature, impact evaluations are intended to answer high-stakes questions. Often (but not always) they are lengthy, complicated, and expensive. So what circumstances would call for doing one?

Well, at a minimum one would want the program (a) to have a design with a cause-and-effect logic connecting activities and outcomes; (b) to have reached a level of maturity with the capacity to offer its services reliably, at a high level of quality, to a significant number of service recipients, and to be able to recruit enough candidates for enrollment that a counterfactual group can be recruited and sustained; (c) to be able to maintain high levels of participation and low levels of premature drop-out; and (d) to be embedded in a secure and sustainable institutional setting.[72]

In addition, the program's intended results should be socially significant – important for the lives and prospects of participants. It would not be responsible to enroll large numbers of people in big programs without knowing to what extent they are likely to benefit meaningfully from the outcomes the program promises them. As a case I point, a few years ago I was asked to review a British six-month long literacy program for children who, by the third grade, had fallen behind their peers in reading; and it offered them a remedial program taught by certified teachers. A full RCT evaluation with thousands of participants had found a statistically significant program impact: after six months participants gained an average of seven points on a ten-point literacy scale. But it turned out that the scale was one that the evaluators had designed, it was not used in schools, and according to the teachers who had done the tutoring, a seven point improvement on the scale actually was so small when looked at in terms of what it meant for improving students' functional reading ability that their actual benefit from the program was meaningless. **So: statistical significance in itself is meaningless unless it pertains to something that is socially significant.**

Statistical significance in itself is meaningless unless it pertains to something that is socially significant – important for the lives and prospects of participants.

Not surprisingly, the U.S. Departments of Education, Labor, and Criminal Justice all support impact evaluations and maintain databases that provide various answers to the question: What Works? [73, 74]

Another situation in which impact evaluations should be used is when plans are made to "scale up" or replicate small to mid-sized programs with the intent of bringing them to hundreds and even thousands of people. It is a matter of ethics and social responsibility.

Here it is worth mentioning the exemplary case of the Nurse-Family Partnership (NFP), a program for first-time pregnant young women in which specially trained APRN nurses begin regular home visits early in the mother's pregnancy and continue them until the child's second birthday. They use checklists to make sure that they pay attention to the physical safety and emotional wellbeing of family members every visit, including checking on conditions of the living space that might pose hazards to a young child. David Olds, a psychologist who developed this program, first tested it in an RCT in 1977 in Elmira, NY with a population that was mostly semi-rural and white. This study showed positive impacts on child health and safety, as well as on the mothers' decreased mortality rates related to pregnancy and birth. Further, the mothers showed decreased pregnancy-induced hypertension and reduction in the likelihood of a next pregnancy within 6 months of childbirth. The young children also benefitted significantly: among other things, there was a 48 percent reduction of child abuse

[72] Hunter, D. E, K. (2006). Op. cit. pp. 180-185.
[73] These include the Department of Education's "What Works Clearinghouse" at https://ies.ed.gov/ncee/wwc/; the Department of Justices' "Virtual Library and Abstracts Database" at https://www.ojp.gov/ncjrs/virtual-library; the Department of Labor's "Clearinghouse for Labor Evaluation and Research" at https://clear.dol.gov/.
[74] However, actual government evaluation practices fall far below what is intended. An assessment by the Government Accounting Office showed that "...most federal managers lack recent evaluations of their programs. Forty percent reported that an evaluation had been completed within the past 5 years of any program, operation, or project they were involved in. Another 39 percent of managers reported that they did not know if an evaluation had been completed, and 18 percent reported having none. See: Program Evaluation. (2017). United States Government Accounting Office Report to Congressional Committees.

and neglect and a 67 percent decline of behavioral and intellectual problems at age 6.[75]

But David Olds was not convinced this program would work in different contexts or with different ethnic groups. So in 1990 he launched a second RCT in Memphis, TN with participants who principally were African American, and found very similar impacts. Then, four years later in 1994 he launched the third RCT in Denver, CO with predominantly Hispanic participants, and again the impacts were similar. **It was only after these trials that Olds was confident enough in the program's effectiveness that he could agree to its replication across the country – which began in earnest in 2003 with the establishment of the Nurse-Family Partnership National Service Office in Denver, where all nurse practitioners in the program are required to receive their specialized training.** Currently, NFP has been replicated in 40 states and has served more than 360,000 families. It is also replicating in England, Australia, Northern Ireland, Canada, Scotland, Norway, and Bulgaria.[76] The scale-up has been very successful and is supported by hundreds of millions of dollars in private funds as well as public contracts. But none of this would have happened if David Olds hadn't insisted on proving program impacts in various contexts and with various groups and used RCT evaluations to do so – thereby eliminating selection bias, maturation bias, and history bias.

Finally, although this rarely happens, once a program has been scaled up or replicated based on a positive impact evaluation, it would be a very good idea to conduct another such evaluation across all newly established sites because we know that replication can lead to profound diminishments in program effectiveness. This is far from an idle assertion. Consider the case of the Center for Employment Training, originally located in San Jose, CA. Its program is to train young people in transitional employment settings and place them with partnering businesses, and it…

> …had shown great promise in the 1980s with large positive effects on their employment and earnings…. Based on these earlier results, the U.S. Department of Labor launched the Evaluation of the Center for Employment Training Replication Sites in the mid-1990s, which was designed to test whether the CET model could be implemented successfully in different settings and have similarly positive effects on the youth served. This final report on the evaluation summarizes the replication effort's success and effects on youth after four and a half years. It shows that, even in the sites that best implemented the model, CET had no overall employment and earnings effects for youth in the program, even though it increased participants' hours of training and receipt of credentials.[77]

Gordon Berlin, MDRC's then-President commented that "…the findings do raise questions about whether a dynamic program like CET can, in fact, be replicated. CET-San Jose is unique in so many ways, having grown organically over 20 years, with an unusually committed founder and staff, very strong ties to the local community, and a tradition of political advocacy on behalf of the local Hispanic community. Perhaps a homegrown model like CET cannot be easily exported in a top-down way to other areas. More research is needed on how to transfer promising models to other areas, particularly given the difficulties that at-risk youth face in today's competitive job market." [78]

Finally, many evaluators believe that it makes little sense to conduct an impact evaluation with a program that has not been evaluated with regard to how well the program model or design has been implemented. We discuss this matter below under the heading of implementation evaluations.

And here is an important point not made often enough about **the value of an evaluation that finds no impacts:**

---

[75] NFP is among the most frequently evaluated programs in the country – and probably in the entire world. Some of the studies showing these impact data are: 1. Olds, D. L., Kitzman, H., Knudtson, M. D., Anson, E., Smith, J. A., & Cole, R. (2014). Effect of Home Visiting by Nurses on Maternal and Child Mortality: Results of a 2-Decade Follow-up of a Randomized Clinical trial. JAMA Pediatrics 169(9). pp. 800-806. 2. Kitzman, H., Olds, D. L., Henderson Jr., C. R., Hanks, C., Cole, R., Tatelbaum R., McConnachie, K. M., Sidora, K., Luckey, D. W., Shaver, D., Engelhardt, K., James, D., & Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. A randomized controlled trial. JAMA 278(8). pp. 644-652. 3. Olds, D., Eckenroad, J., Henderson, C. R., Kitzman, H., Powers, J., Cole, R., Sidora, K, Morris, P., Pettitt, L., m. & Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. Fifteen-year follow-up of a randomized trial. JAMA 278(8). pp. 637-643. 4. Karoly, L. a., Kilburn, M. R., & Cannon, J. S. (2005). Rand Corporation. 5. MacMillan, H., Wathan, L., Barlow, N. C., Fergusson, J., Leventhal, D. M., Taussig, J, M., & Heather, N. (Interventions to Prevent Child Maltreatment and Associated Impairment. Lancet. pp. 1-17.

[76] International replication of the NFP has not been executed as rigorously. In Australia, the program has begun enrolling selected mothers with previous children, a very significant departure from the U.S. model. Bulgaria is undertaking an implementation evaluation; Canada has launched an RCT evaluation, but only after having begun a significant amount of replication; and England completed an RCT in 2015 – but when I was invited to review the program's implementation there in 2005 I found significant systemic problems with the level of control that the central office was able to exert over fidelity of implementation across sites to the program's design. To my knowledge evaluations have not be used to support the implementation of the NFP in the other countries. NFP's international replication is discussed at: https://nfpinternational.ucdenver.edu/international-program

[77] Miller, C., Bos, J. M., Porter, K. E., Tseng, F. M., & Abe, Y. (2005). The Challenge of Repeating Success in a Changing World: Final Report on the Center for Employment Training Replication Sites. MDRC:xi.

[78] Berlin, G. Ibid.

> It would be a mistake…to write off a program entirely if an impact evaluation reveals that it is not producing impacts as intended. Nor is it sensible to try to cover this over. "Often, an overall null effect in a trial leads to authors cherry-picking any favourable results and giving them undue prominence, or service implementers querying the veracity of the methods used, such as study size or choice of measures, suggesting these are responsible for the failure to detect a positive impact. There is also a tendency for the academic field to be less interested in null effects – manifested, often, in a failure to publish results. As one paper on this topic put it, [we seek]…to model a more positive and thoughtful response to finding null effects. Far from a null effect from one high-quality trial necessarily equating to a failed service, we believe that it can point to valuable learning; indeed, if the results are used and acted upon, it can be part of a normal and healthy process of service improvement.[79]

Finally, here are some points worth keeping in mind when considering what approach to use in an impact evaluation.[80] The evaluation design should:

- Be able to identify multiple causal factors,
- Make it clear how causal effects will be examined and how claims of causality will be developed,
- Be able to identify how the program works and why, and
- Identify possible alternative explanations for program results and eliminate them.

Also, the evaluators should:

- Make their values and special interests clear from the start of the engagement,[81] and
- Make very clear to participants in the evaluation what risks or costs it might pose for them and how these risks will be meliorated.[82]

It is also worth mentioning that **the strength of causal inference is inversely correlated with the scope and complexity of its delivery.** This becomes very relevant for programs that are national in scope such as, for instance, Communities in Schools (which supports at-risk students) or the Youth Villages Intercept program (which focuses on children in foster care or who have aged out of the system); both have been evaluated and found to have positive impacts – but Youth Villages' program has a much more tightly defined target population and a very focused and intensive model; therefore, the causal findings in the case of Youth Villagers may well be stronger than for Communities in Schools. "Designs and methods applied to narrowly specified interventions can support strong causal claims but as the scope and scale of an intervention increases the strength of causal claims is reduced. The reality is that many contemporary programmes are not narrowly specified: they are ambitious, broad in scope and made up of many subprogrammes. Policy makers may therefore have to accept a trade-off between strong causal inference and relevance." [83]

> The inability of a program to show impacts in a given high-quality trial does not mean that the enterprise is a failure. We believe that it can point to valuable learning, and, if the results are used and acted upon, it can be part of a normal and healthy process of service improvement.

What are the usual challenges in conducting an impact evaluation?

Generally speaking, any program will encounter

[79] Axford, N., Whelan, S., & Hobbs, T. (2015). Wrong Answers, Right Response: Learning from randomised control trials when you don't get the results you were hoping for. Realizing Ambition – Program Insights: Issue II. Big Lottery Fund. p. 5. [Author's note: The Big Lottery Fund, supported by revenues from Britain's national lottery, has been renamed and now is called The National Lottery Community Fund.]
[80] Stern, E., Stame N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the Range of Designs and Methods for Impact Evaluation: Report of a Study Commissioned by the Department for International Development. Department for International Development (U.K.). Working Paper 38.
[81] For instance, academic evaluators at times will have their own research agendas and add to the cost of an evaluation by insisting on piggybacking extra metrics onto the data gathering processes.
[82] For example, the fact that if they are in the control group access to the program will be denied to them for the duration of the evaluation.
[83] Stern, S., Stame, N. et al. (2012). Op. Cit. p. 80.

significant challenges when conducting an impact evaluation. These include:

- **The challenge of the counterfactual:** A main issue in any impact evaluation, as discussed above – namely, how to design and obtain reliable information from the counterfactual to which the program being evaluated will be compared.

- **Problems of ethics:** This is the question that arises in selecting a counterfactual that most often troubles practitioners – that is, whether or how to create a counterfactual or comparison group that involves denying services to individuals who otherwise would be eligible to participate in the program.

- **The challenge of fidelity:** Once a program has decided to undergo an impact evaluation, it must hold constant for the duration of the evaluation and also be the same and hold constant across all sites. Shifts in the profile of program enrollees (demographic and risk indicators), in program elements, in dosages, in staff training, in supervision methods, in program quality standards, and even in the contexts in which the program is offered can have important consequences for program outcomes. Where such shifts happen they make it impossible to know what exactly one is evaluating. This is often a big sticking point for program staff who are very alert to issues of ethics described above.

- **The challenge of operating a "frozen" program:** The requirement to refrain from making program adjustments or adding new elements for the course of the study, which often will last several years. The longer the program, the longer the period an evaluation will have to cover. This can produce significant stress for front-line staff and their supervisors, especially when, in the course of their work, they are discovering programmatic gaps or deficiencies.

- **The issue of cost:** Driven by the need for evaluators to make repeated measurements and assessments with individuals in the program being studied and those in the counterfactual – and doing so over a multi-year period – as well as analyzing the data and then facilitating meetings with staff to arrive at meaningful and useful interpretations of the data.[84]

- **The further issue of transaction costs:** The inevitable burdening of program staff with evaluative tasks that have no apparent value to them in their work but are necessary for the study.[85] It is essential that staff understand the evaluation's goals and have a role in making key decisions about how it will proceed. The evaluation's planners should take pains to help staff understand that the evaluation will, in the long run, be of benefit to them because it will provide essential information to help evolve the program further, make it more effective, and possibly extend it to many more participants. Without staff buy-in, the evaluation will be weakened considerably since any robust evaluation will include a variety of activities in which staff will be required to participate, such as the following:

  - Focus groups to explore qualitative issues and implementation challenges,

  - Arranging and managing focus groups for program participants to illuminate issues of program quality,

  - Taking on some or even all of the evaluator's measurement requirements,

  - Managing the usual work flow while arranging for various meetings and groups required by the evaluator,

  - Advocating for changes in the evaluation plan in response to program participants' needs and desires, and

  - Participating in data analysis meetings to provide staff perspectives.

Impact evaluations, it should be clear, are not right for all programs at all times, and alternative evaluation methodologies are discussed below. But before taking on the challenge of an impact evaluation on any kind – using an RCT, quasi- experimental methods, or even benchmarking – it is essential that the question of what is being evaluated be looked at deeply and answered clearly. Why? Because if this is not well understood, the evaluation will pretty much be useless; or worse, it might seem to prove something which patently is untrue. Answering the question of "the what" is best accomplished through an implementation evaluation.

[84] Cost is driven by the length of follow up needed to learn whether the outcome of interest changed – for example, it is one thing to learn whether a program increased participants' attainment of GEDs, another altogether to learn whether that led in turn to job placement in career ladder jobs or enrollment in and completion of post-secondary education. Cost also is determined by whether or not the measure requires a survey or can be gleaned much less expensively from administrative records data as discussed below.

[85] Patrizi, P. & McMullen, B. (1999). Realizing the Potential of Program Evaluation. Foundation News and Commentary. 40(3). pp. 30-35.

Function 3: Assessing Program Fidelity of Implementation

In this section we look at how well a program has been implemented when compared to its original design – that is, the fidelity of implementation. But what is fidelity? As in a marriage, program fidelity involves holding true to foundational commitments and key ingredients that, In this section we look at how well a program has been implemented when compared to its original design – that is, the fidelity of implementation. But what is fidelity? As in a marriage, program fidelity involves holding true to foundational commitments and key ingredients that, arguably, make both marriages and programs worthwhile. Speaking more technically, "Fidelity may be defined as the extent to which delivery of an intervention adheres to the protocol or program model originally developed." [87]

This becomes an especially important question when evidence-based programs are to be replicated or scaled up. But what is an evidence-based program? The answer requires two parts. The first has to do with evidence-based practices, which are identified by research and provided to people requiring treatment or other supports in order promote the best possible results for their recipients. The second has to do with evidence-based programs, which are codifications of grouped evidence-based practices that research shows are linked to expected outcomes within the organizational systems, processes and supports necessary for their delivery.[88]

So: In the context of continued pressure to scale up programs, the question of whether a program deserves to be labeled evidence-based becomes a matter of concern, and this has led to a focus on implementation evaluations.[89, 90]

Currently CTOP is working in partnership with the City of Hartford, the Hartford Foundation for Public Giving, and the Tow Foundation to support Roca, headquartered in Chelsea, MA, in replicating its Young Mothers Program in Hartford, CT. This program engages young mothers who, by virtue of their association with gangs and/or lack of connection to any prosocial institutions, are at high risk for committing or being victims of violence and other overwhelming experiences with the consequence that the prospects for them and their children of escaping poverty are low to none. This program is trauma-informed and incorporates several other evidence-informed practices, and has demonstrated its effectiveness in two benchmarking studies.[91, 92] A local needs assessment

[87] Mowbray, C. T., Holter, M. C., Teague, T. B. & Bybee, D. (2003). Fidelity Criteria: Development, Measurement, and Validation. America Journal of Evaluation (24). p. 315.
[88] Metz, A. J. R., Espiritu, R. & Moore, K A. (2007). What is Evidence-Based practice? Research to Results Brief. Child Trends.
[89] There is an extensive literature on what has become known as "implementation science". See, for example: Fixsen, D. L., Naoom, S. F., Blaise, K. A., Friedman, R. M., & Wallace, F. (2005). Implementation Research: A Synthesis of the Literature. Louis de la Parte Florida Mental health Institute, University of Florida.
[90] Often, the terms "implementation evaluation" and "formative evaluation" are used interchangeably. However, the term "formative evaluation" (first introduced by Michael Scriven along with "summative evaluation") also is used in a very specific way to refer to evaluations of programs that have reached the half-way point of their intended duration and to look at whether they are producing outcomes as expected. In this framework, at the end of a program it is then subjected to a "summative evaluation" that is intended to test its impacts. CTOP generally uses the terms "implementation evaluation" and "impact evaluation" in their more general senses to avoid possible misunderstandings.
[91] Crime and Justice Institute. (2012). Benchmarking: How does Roca's High-Risk Intervention Model compare to other high risk youth Programs? Crime and Justice Institute.
[92] Godley, S. (2017). Improving Outcomes for Teen Parents and their Children in Massachusetts 2017: An Analysis of Population Changes and Service Needs. Unpublished Doctoral Dissertation at Boston University School of public health.

found that no other social service entity in the Harford area is focused on this group of young women and that their requirements for services are extremely high. So it made a great deal of sense to key stakeholders for Roca to replicate this program in that city.

However, just because the program has demonstrated its effectiveness in other venues, there is no reason to assume that it will do so when fully implemented in Hartford. Why? Because it is well known that replication is a risky business as far as maintaining program effectiveness is concerned – because the context is new, the institutions with which the program must engage to support its clients are new to it, and the organizational culture, systems, and processes that supported the program's effectiveness in prior venues may not be fully realized in this new setting. Therefore, Roca and CTOP are planning for an implementation evaluation of the Hartford site for late in 2022 so that all stakeholders can be confident that the young women will be benefitting as expected. And some years after that, if all goes well, it will be time to do an impact evaluation.

But while the case for focusing on fidelity is sound, it is far from uncomplicated, as a study by Bridgespan revealed:[93]

> La Alianza Hispana, a Roxbury, Massachusetts-based nonprofit serving the Latino community, chose Cuidate as its EBP because the agency was already delivering the six-session intervention and had experienced facilitators on staff. As implementation began, this presumed asset began to melt away. "Before we got the grant, we had five facilitators doing Cuidate, and each was doing it their own way," explained Program Director Lily Rivera. "People were picking out and doing the things that they liked. The…performance measures, observation requirements, and evaluations all have led us to have much more fidelity in how we do this. We no longer skip session six because two kids once got into a fight."
>
> But doing Cuidate by the book was not easy for the experienced facilitators. "We had some very educated facilitators who couldn't get their heads around this," Rivera said. "Staff were coming and going. But we had to get the right staff. I ended up going to a probationary period of 90 days, so I could confirm that the people I hired and trained would follow through."

As a precursor to replicating a program, at least three questions should have been asked and answered affirmatively:[94]

1. Do we have evidence that our program produces positive results?
2. Do we know which elements of our program are required to be effective?
3. Are our current organization and finances strong? [95]

If the answer to each of these questions is not an unambiguous yes, replicating in the hopes of developing social value (see Box 4) is a fool's errand.

Box 4 How can social value be measured? [96, 97]

Some metrics commonly used to measure social value:

1. **Cost-benefit analysis**
2. **Cost-effectiveness analysis**
3. **Impacts proven via evaluation**
4. **Calculations of monetized social return on investment**
5. **Measures of public support for, or valuing of, program results**
6. **Life satisfaction assessments**
7. **Calculations of added life years**

[93] Stid, D., Neuhoff, A., Burkhauser, L., & Seemann, B. (2013). What Does It Take to Implement Evidence-Based Practices? Bridgespan.
[94] Campbell, K., Taft-Pearman, M. & Lee, M. (2008). Getting Replication Right: The Decisions that Matter Most. Bridgespan.
[95] While this question goes beyond the program itself, if the organization isn't strong and financially secure it will not be able to create the new organizational structures and capacities to deliver the program effectively, reliably, and sustainably in new venues or contexts.
[96] Mulgan, G. (2010). Measuring Social Value. Bridgespan.
[97] A thoughtful article on measuring social value is Clifford, J., Markey, K, & Malpani N. (2013). Measuring Social Impact in Social Enterprise: The state of thought and practice in the UK. E3M.

When replicating a program, it is important that key implementation or replication standards be established and communicated to all parties involved. While this in itself is a complex matter, [98] the following items should be identified and specified in advance of any replication effort: [99]

- **The core program elements** that must be replicated along with research-based understanding of their efficacy and effectiveness; [100]
- **The key implementation components,** that is, such things as necessary staff competencies, what staff supervision methods will be used, essential administrative structures and systems needed to support program delivery with fidelity;
- **Program elements that can be modified** in the course of replication, and those that can be omitted altogether; and
- **The nature of technical assistance and consultation** that will be provided to those charged with the replication.

Before moving on to discuss implementation evaluations, it is important to understand the inherent tension that program managers face between the need to optimize program performance by setting guardrails, performance standards, and staff expectations on the one hand – and on the other hand empowering staff to work creatively or even innovatively in providing their services (knowing, for example, when they can take initiative to modify something they are doing in response to emerging circumstances). There is no more sure-fire way of driving down staff morale than by attempting to manage them in a top-down, command and control manner. And low staff morale certainly will not engender high performance.

To summarize: Rather than establishing whether or not programs work, implementation evaluations look at **how programs work and whether they are working as designed.** This entails looking at and then beyond the program at the organizational characteristics – including its culture, resources, systems, and processes – that determine to a large degree how well, how reliably, and how sustainably a program is delivered as designed.

Unlike an impact evaluation that looks backward and addresses the question of whether a program has produced the results for which it was intended, an implementation evaluation is forward-looking and is concerned with program improvement. **The intent is to support improvement, not establish a judgement.** [101]

A well-executed implementation evaluation, which by its nature is designed to reveal in what ways a program is working well or as expected – and in what ways it isn't – should lead to significant stock-taking by the organization and the development of a plan, when this is called for, to improve, build out, or build up its competencies and capacities, systems and processes, and programming. Generally speaking, prior to undertaking an impact evaluation, programs are well advised first to undertake at least one implementation evaluation and then make necessary tactical and strategic adjustments and adaptations.[102]

> Improvement-oriented evaluations ask the following kinds of questions: What are the program's strengths and weaknesses? To what extent are participants progressing toward the desired outcomes? Which types of participants are making good progress and which types aren't doing so well? What kinds of implementation problems have emerged and how are they being addressed? What's happening that wasn't expected? How are staff and clients interacting? What are staff and participant perceptions of the program? What do they like? Dislike? Want to change? What are perceptions of the program's culture and climate? How are funds being used compared with initial expectations? How is the program's external environment affecting internal operations? What efficiencies can be realized? What new ideas are emerging that can be tried out and tested? [103]

Another use of implementation evaluations is to monitor how well a program is being replicated at new sites (especially if it is evidence-based).

---

[98] Kusek, C. J. & Rist, R. C. (2004). Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners. The World Bank.
[99] Metz, A. J. R., Bowie, L. & Blasé, K. (2007). Seven Activities for Enhancing the Replicability of Evidence-Based Practices. Research to Results Brief. Child Trends.
[100] Efficacy refers to how well a program produces intended results during its pilot stage or when it has been implemented under controlled conditions, with very tight standards, and has continuously been monitored for fidelity compliance; effectiveness refers to how well a program produces intended results in its day to day operations in a less controlled context.
[101] Bowie L. & Bronte-Tinkew, J. (2008). Process Evaluations: A Guide for Out-of-School Time Practitioners. Research to Results Brief. Child Trends.
[102] While some evaluators build implementation evaluations into impact evaluations, in my view this is unfortunate because it gives the organization little or no time to make the adjustments or improvements flagged as needed through the implementation evaluation – and therefore the program is less likely to be as effective at producing impacts than it might have been had it been given time to make needed changes.
[103] Patton, M. Q. (2008). Op. cit. (pp. 116-117).

An implementation evaluation will typically involve the following elements:

- Description of the context within which the organization works and the need for its programming/services;
- Review of the theory of change and the degree to which it is informed by evidence-based practices;
- Assessment of the fidelity of programming/services as they actually are being delivered compared to what is called for in the theory of change;
- Assessment of the quality of program and service delivery; [104]
- Review of the methods used to recruit and enroll participants so that conformance with the target population as defined in the theory of change is assured;
- Regression analysis of programming/service components that seem to drive participants' progress toward achieving short-term and intermediate outcomes as intended;
- Description of staff competencies compared with the competency profile needed to deliver high quality and effective programming and services to the target population as designed;
- Analysis of program participants' demographic and risk characteristics compared to those called for in the theory of change;
- Analysis of participants' program completion rates in aggregate as well as disaggregated in terms of demographic and risk-level characteristics;
- Description of the short-term outcome "ladders" used to help participants achieve the program's intermediate outcomes;
- Description of the organization's intermediate outcomes that mark significant progress toward participants achieving long-term outcomes as intended;
- Analysis of the rates participants achieve short-term, intermediate, and long-term outcomes – aggregated and disaggregated in terms of demographic and risk-level characteristics;

The intent of an implementation evaluation is to support program improvement, not establish a judgement. When well-executed, it should lead to significant stock-taking by the organization and the development of a plan, when one is called for, to improve, build out, or build up its competencies and capacities, systems and processes, and programming.

- Analysis of the social significance of the outcomes the organization is working to achieve with its program participants;
- Audit of the organization's internal performance data (looking at their accuracy, validity, reliability, and also timeliness of the entry of these data into its performance management system);
- Description of how internal performance data are used (or not used) to undergird strategic leadership decisions, as well as day-to-day management decisions and decisions made by front-line staff in their everyday delivery of programming/services;
- Identification of possible or actual legal exposure(s) faced by the organization.

An implementation evaluation is often just as complex as an impact evaluation, and requires a considerable amount of time and effort – and cost.

And now we turn from what some might call "evaluation proper" – that is, evaluation as historically conceived of by professional evaluators – to the matter of how evaluations and evaluative thinking can be used to support program development, implementation, improvement, learning, and performance management.

[104] Generally these are assessed through direct observation, but also via focus groups or surveys of participants, front-line staff, and their supervisors.

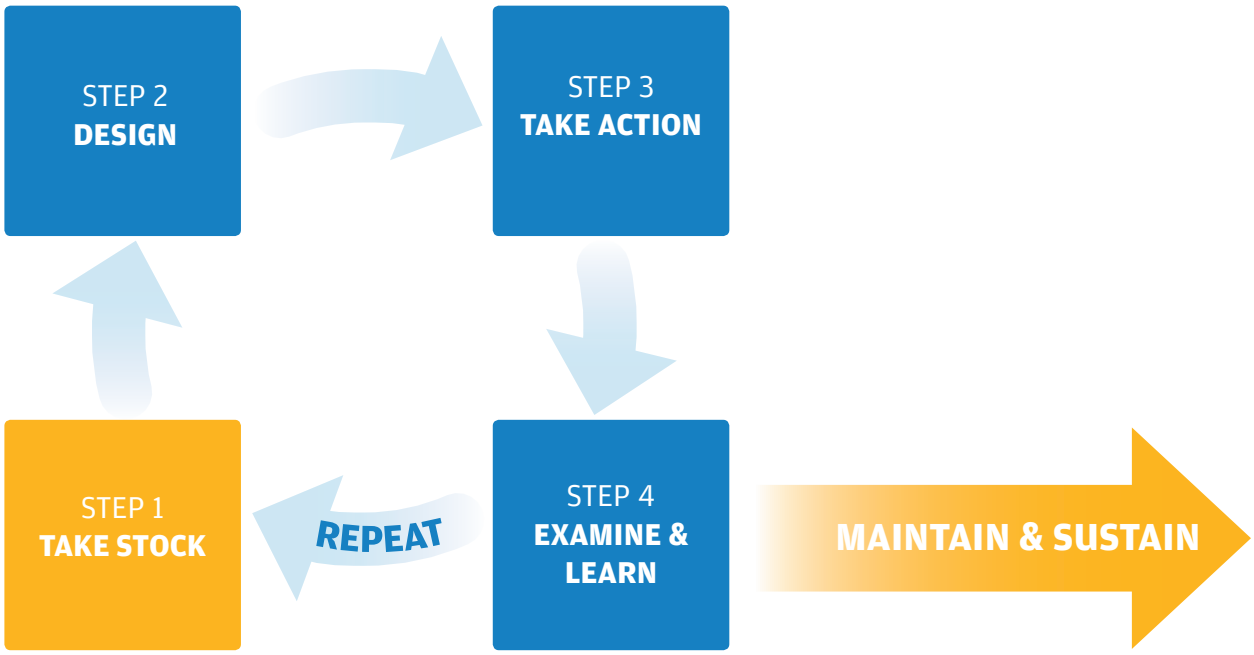
Function 4: Informing Program Development, Implementation, Improvement, and Management

In this section we discuss evaluative methods that have emerged over the last two decades. Under the general rubric of "rapid cycle evaluation" these approaches are about getting results in a relatively short period of time, providing feedback loops that can be used to make program improvement. **A key difference between rapid cycle evaluations and implementation evaluations is that the latter look at whole programs, while rapid cycle evaluations look at individual program elements (e.g., case management) or processes (e.g. managing program enrollment) in order to help improve them.**

> [Rapid cycle evaluation] approaches use rapid cycle analysis as part of a feedback loop to generate new knowledge and optimize interventions. These approaches enable the engagement of stakeholders—the people and agencies invested in the program and interested in the evaluation results—in reviewing program components, analyzing and interpreting results, and adapting practice and measurement collaboratively. Stakeholders have key insights into the program, which provides the information required to quickly tailor design components to the local context.[105]

As might be expected, evaluators have developed various models for conducting rapid cycle evaluations, but all of them have in common a circular set of elements that are very similar from one model to the next (see Box 5), although they often have different names.

Box 5 Generic elements of rapid cycle evaluation



[105] MDRC. (2020). Rapid Cycle Evaluation. Research Brief.

Step 1: Take Stock - which entails asking questions such as, what are the major issues of concern to the people with whom we are working? How have we been addressing them? How successful have we been? Is there any way in which we should change what we are doing? Then, choosing a specific program element or delivery process on which to focus.

Step 2: Design - which entails planning changes in the area that has been selected for improvement. This is best done with input from key stakeholders.

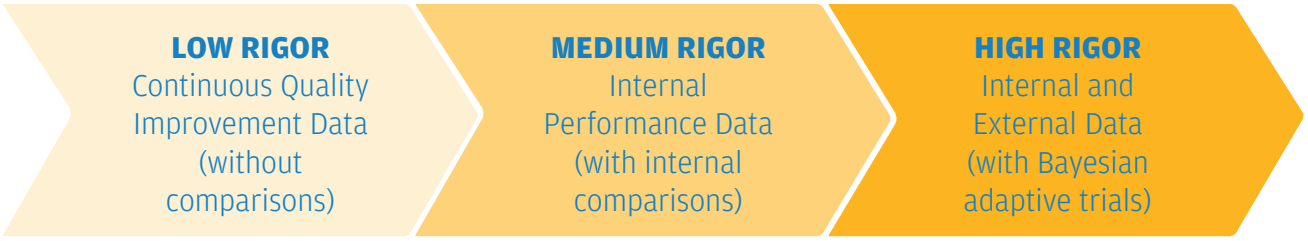
Step 3: Take Action - Implement and monitor - which is the point where program staff put the new designs to work (make them operational) and the evaluation team provides consultation and support both on implementation and on data collection.

Step 4: Examine and Learn - which is the step where, at an agreed upon date, data collection is suspended and the data are mined to learn about how the work has proceeded, where it has succeeded, where it has fallen short, and why. The idea here is not to look at outcomes, on average, for program participants, but rather to understand how well its efforts are adapted to the context, what its strengths are, what its weaknesses are, where its opportunities lie, and what threats it faces (which actually amount to the components of a traditional SWOT analysis[106]). And here, then, is where the four questions of performance management outlined in Box 1 at the very beginning of this article come into play.

And then the cycle starts all over again, until key issues have been resolved, the questions answered, the program suitably adjusted, and it is ready to be maintained at high levels of quality and to be sustained until a new cycle is required or subsequent implementation and impact evaluations can be undertaken when the time is right.

How are rapid cycle evaluations designed? It's useful to think in terms of a continuum of rigor, as shown in Box 6 and discussed on the next page.

Box 6 A continuum of rigor for rapid cycle evaluations[107]



[106] A useful discussion of SWOT analysis can be found at: https://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/swot-analysis/main
[107] The illustration and discussion are from Ibid. pp. 3-4.

Low rigor rapid cycle evaluations are used to support program implementation and for continuous quality improvement (CQI). Program monitoring is ongoing, data are collected continuously – and changes are assessed periodically. Sometimes referred to as "Plan-Do-Study-Act" cycles, these kinds of evaluations are best used when a program is trying to improve on aspects or elements of what it is doing, rather than as a way to determine the effectiveness of a new element or innovation.

Medium rigor rapid cycle evaluations use quasi-experimental methods that establish previously existing internal performance data against which a program change is being measured. Strict guidelines are used for maintaining data quality, and recognized statistical methods are used to assess changes. The evaluation design may simply involve "before and after" assessments, but more sophisticated methods also are used such as "interrupted timeseries" where a sequence of implementation steps result in a new baseline and progress is measured progressively against how these baselines change. Often regression analyses are applied; these use statistical analyses to establish linkages between the changes that were introduced and what happened subsequently.

High rigor rapid cycle evaluations make use of Bayesian Trials, which combine as many external data sources as possible – including any available qualitative reports, opinions expressed by experts or practitioners, as well as any quantitative studies – to establish the "common sense" likelihood that the data showing emerging variations in program performance indeed were caused by the changes that were made. About Bayesian Trials (in this case discussing impact evaluations but making the point that applies here as well):

> Consider two identically designed studies of an employment intervention. Data from the first study show an increase in employment of 6 percentage points while data from the second one show an increase of 5 percentage points. A classical researcher notes that the p-value of the first estimate is 0.05, so the estimated effect is statistically significant, while the p-value of the second is 0.11, indicating the estimate is not statistically significant. In the classical world, the first result would typically receive much more attention than the second even though they differ by only one percentage point. A Bayesian analysis using a weakly informative prior would, by contrast, indicate there is an [sic] 94.5 percent probability that the impact is positive in the second study and a 97.5 percent probability that the impact is positive in the first study. The Bayesian analysis would thus favor the first finding, but the difference between them would be presented as relatively small, as seems reasonable when the estimates differ by only one percentage point.[108]

Whenever possible, Bayesian Trials should be used in combination with external counterfactuals be they benchmark data, comparison or control groups.

As already mentioned, **in contrast with impact or implementation evaluations that look at programs as wholes, rapid cycle evaluations focus narrowly on a specific aspect of a program or a specific idea for improving the program that its providers want to try out before adopting it as part of the full program design.**[109]

While well-executed rapid cycle evaluations attempt to be as rigorous in terms of comparisons and control groups as feasible, less meticulous methods often are adopted since the nature of these evaluations usually means smaller sample sizes and thus less confidence in the statistical significance of findings. So a caveat is in order here: **A common mistake is for people to think that a rapid cycle evaluation is an impact study proving the program works.**[110]

In other words, rapid cycle evaluation is utilization-focused.[111] For pragmatic purposes such evaluations need to be highly tailored to local circumstances, to the unique qualities of the program under consideration, and to the nature of the question(s) being asked. And, of course, they have to greatly shorten the usual sequence of data collection, data analysis, data reporting, and then implementation of the program adjustments that they inform; and then, as needed, the whole cycle might start all over again. Hence the term "rapid cycle evaluations", indicating they rely on rapid evaluation methods.[112]

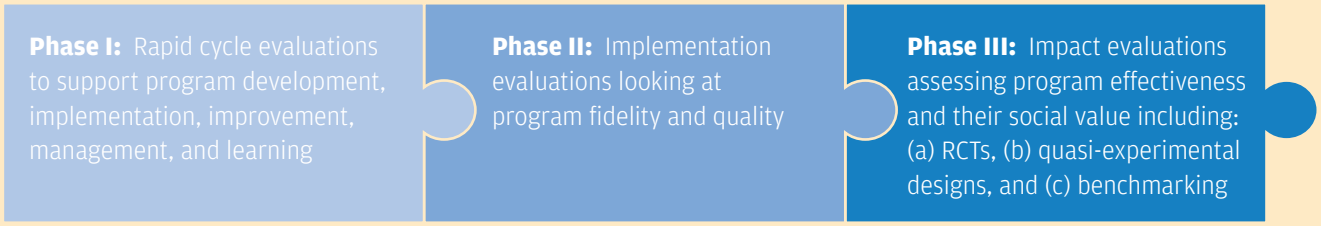
---

[108] Michalopoulos, C. (2018). Bayesian Methods in Social policy Evaluations. Reflections on Methodology. MDRC. p1.
[109] MDRC. (2020). Op. cit.
[110] I am grateful to Gordon Berlin (2021 Op. cit.) for making this point to me.
[111] Patton, M. Q. (1979). Op. cit.
[112] It should not be surprising that many professional evaluators resist this approach since it challenges both their "hands off" neutral stance as well as the usual pacing of their work when doing impact, benchmarking, and implementation evaluations.

They are iterative and focused, examining selected aspects of program planning, design, implementation, adaptation and improvement on which an organization is focusing at a given time – and for a limited purpose – in its developmental arc. Thus, **how** an organization or program makes use of this kind of evaluation is as much a part of the evaluative effort as the process itself – another characteristic that distinguishes rapid cycle evaluation from more traditional evaluation methods.

Box 7 A developmental sequence for using evaluation to drive program performance

In this paper we started with a discussion of impact evaluations, then moved on to benchmarking evaluations that demonstrate program effectiveness. Next we considered implementation evaluations. Finally we turned to evaluations that are appropriate to developing, implementing, improving, managing, and learning from programs. Thus, so far in this paper we are working backwards. But a good way to look at evaluations developmentally would be:



Finally, rapid cycle evaluations require a robust theory of change (discussed earlier in this paper) as a conceptual framework within which to work. Where one has not yet been developed, evaluators should regard helping in its creation as a first step prior to launching the evaluative cycle. Of the various approaches to rapid cycle evaluation, CTOP has decided to make use of a particular one that seems well designed to support our organizational improvement and growth. It is called developmental evaluation, and the following discussion should help the reader to understand why we have decided to make use of it.

Developmental evaluation

Michael Quinn Patton, a qualitative[113] and utilization-focused[114] evaluator, is also the father of developmental evaluation – he published the first paper on the topic in 1992 and in 2011[115] published what now has become a foundational volume on the topic.

What is developmental evaluation?

One way of answering this question is to compare developmental evaluation with implementation evaluation. Whereas the latter focuses on the matter of **program improvement** (often as a prelude to an impact evaluation), developmental evaluation focuses on **program adaptation** – to changes in the environment or context as well as changes in the population being served, but also in response to what has been learned through the course of operating the program; or, to the fact that new methods have emerged for addressing the issue that the program was intended to address in the first place. In a sense, developmental evaluation provides a way for reframing, dynamically, what the program's purpose is and what it must do to adapt to the context(s) within which it is working and to the people whom it is serving. This means that, **in contrast to implementation evaluations, developmental evaluations adopt diverse frameworks of inquiry and analysis as appropriate to the situation being studied** – including triangulation (comparing different perspectives on an issue), appreciative inquiry, reflective practice, outcome

[113] Patton, M. O. (2001). Qualitative Research and Evaluation Methods. 3rd edition. Sage.
[114] Patton, M. Q. (2008). Op. cit.
[115] Patton, M. Q. (2011). Op. cit.

mapping, systematic risk management, principles-focused evaluation, comparisons of the actual to the ideal, and even the use of so-called "wicked" (presupposition-challenging) questions.

Eight essential principles combine to define developmental evaluation: [116]

1. **Developmental purpose** – it is conducted as a contribution to program development;
2. **Evaluation rigor** – it maintains an intellectual distance and rigor, a commitment to the scientific approach, even while being embedded in the program design and implementation team;
3. **Utilization focus** – it is intended to be immediately useful to program managers and front-line staff who are, looked at in this framework, the innovators who are using the evaluation most directly;
4. **Innovation focus** – it focuses on changes in program design and delivery that emerge in changing contexts;
5. **Complexity perspective** – it examines the presuppositions and experienced needs of intended program beneficiaries, program staff and managers, local contextual factors, funders both public and private, other stakeholders, as well as environmental contexts such as safety issues, the availability of basic resources, etc. – and of key inter-relationships among them; all of which ultimately adds up to looking at the dynamics of systems change;
6. **Systems thinking** – it is grounded in the understanding that social systems (and natural systems too, for that matter) have important impacts on intended program beneficiaries, program staff and managers, and the organizations that are delivering programming; and that these must be taken into account in attempting to learn about what a program is, does, and accomplishes – and how (and why) these change over time;
7. **Co-creation** – it recognizes that to be relevant and useful, a developmental evaluation design must be created by workgroups in which key stakeholders are represented, valued, listened to, and have their needs addressed; and
8. **Timely feedback** – it is only useful when it collects and analyzes data using methods that allow evaluators to share their (always provisional) findings frequently with teams designing, delivering, and managing the program.

As these principles show, developmental evaluation is inherently a special case of utilization focused evaluation, and is to a large extent an exploration of qualitative issues. It consists of the use of evaluative methods when looking at – and supporting – innovation in areas that include "…creating new approaches to solving intractable problems, adapting programs to changing conditions, applying effective principles to new contexts (scaling innovation), catalyzing systems change, and improvising rapid responses in crisis conditions." [117]

But, since developmental evaluation is a kind of rapid cycle evaluation, whatever methods are used it is essential that the findings they produce are fed back promptly into organizational and program-level decision making.[118]

What conditions are well suited for using developmental evaluation? Patton lists seven: [119]

1. Highly emergent and volatile situations [e.g., the current Covid-19 pandemic crisis];
2. Situations that are difficult to plan for or predict because the variables are interdependent and nonlinear;
3. Situations where there are no known solutions to issues, new issues entirely, and/or no certain ways forward;
4. Situations where multiple pathways forward are possible, and thus there is a need for innovation and exploration;
5. Socially complex situations, requiring collaboration among stakeholders from different organizations, systems, and/or sectors;
6. Innovative situations, requiring timely leaning and ongoing development; and
7. Situations with unknown outcomes, so vision and values drive processes.

[116] Patton, M. Q. (2016). The State of the Art and Practice of Developmental Evaluation. Ch. 1, pp. 5-124 in Patton, M. Q., McKegg, K. & Wehipeihana, Nan (eds.).(2016). Op. cit.
[117] Patton, M. Q., McKegg, K. & Wehipeihana, N (eds.). (2016). Developmental Evaluation Exemplars: Principles in Practice. Guilford Press.
[118] Patton, M. Q. (2016). The State of the Art and Practice of Developmental Evaluation. Ch. 1, pp. 5-124 in Patton, M. Q., McKegg, K. & Wehipeihana, Nan (eds.).(2016). Op. cit.
[119] Ibid. p. 14, Exhibit 1.4.

A fundamental requirement of developmental evaluation is that the evaluator must be flexible and, just like the program(s) and organization(s) being evaluated, must adapt to change – especially in a crisis. As Patton puts it,

> **Everything changes in a crisis.** Embrace change, don't resist it. Program goals may appropriately change. Measures of effectiveness may change. Target populations may change. Implementation protocols may change. Outcome measures may change. This means that evaluation designs, data collection, reporting timelines, and criteria will and should change. Intended uses and even intended users may change. Expect change. Facilitate change. Document changes and their implications. That's your job in a crisis, not to go on in a comfortable business-as-usual mindset. There is no business-as-usual now. And if you don't see program adaptation, consider pushing for it by presenting options and introducing scenario thinking at a program level. Take risks, as appropriate, in dealing with and helping others deal with what's unfolding.[120]

Developmental evaluation is designed to address issues of emergence or newly arising phenomena and contextual complexity – for the world is changing rapidly, is rapidly becoming more interconnected, and therefore is rapidly becoming more interrelated.

Other things that set it apart from traditional evaluative methods is the insistence that evaluations must always consider the systems within which, and in interaction with which, a program and organization is working. Also, the fact that its practitioners do not stand outside the program with which they are working – in fact, they are part of the team that is designing, implementing, or managing the program. They facilitate discussions about how to evaluate what is going on with the program, and are part of the discussions that produce decisions about how to handle program developments and challenges. They pose questions from an evaluative perspective, bring to bear evaluative data, and help in the analysis of data required to produce actionable information. And they do all this with the expectation that, over time, these methods and approaches will become internalized, part of the program and the organization's DNA.[121]

It is useful, when considering developmental evaluation, to look at the tensions that inherently arise from its intentions, values, and methods. Patton lists five: [122]

1. **Ownership tension.** DE works best when those engaged feel ownership of the process and can creatively adapt to local contexts. But the organizations within which DE is supported must ensure that the way DE is conducted is consistent with the organization's mission and policies. This is the classic tension between imperatives emanating from headquarters and the need for people in the field to exercise their prerogative in adapting to context.

2. **Inclusion tension.** DE works best with the sustained inclusion, participation, and investment of a broad cross section of stakeholders who are affected by an intervention. Having this cross section can generate conflicts in setting priorities and adapting as change occurs. Determining what stakeholders are involved in DE, in what ways they are involved, and what responsibilities they have can be an ongoing source of tension.

3. **Standardization versus contextualization tension.** Large international organizations operating in many countries and conducting programs in many sectors need standardized procedures to ensure coherence and accountability. But DE thrives on local adaptability and contextual responsiveness.

4. **The long-term/short-term tension.** Problems of poverty, poor education, low employment, and inequality have deep roots and take time to address. Recognition of this fact has led to large-scale, long-term investments and initiatives based on extensive planning. Organizations have set up procedures to manage and evaluate on a long-term basis. DE involves an ongoing series of short-term, real-time adjustments. The tension enters when deciding how to integrate the real-time orientation of and short-term decision-making in DE into the longer-term decision-making, planning, and accountability cycles of large organizations.

[120] Patton, M. Q. (2020). Evaluation Implications of the Coronavirus Global Health Pandemic emergency. Pulled from his blog: Evaluation Implications of the Coronavirus Global Health Pandemic Emergency | Blue Marble Evaluation
[121] Patton, M. Q. (2005). Developmental Evaluation. In Mathison, S. (ed.) (2005). Encyclopedia of Evaluation. Sage. p. 116.
[122] Foreword to Baylor, R. et al. (2019). p. 2.

5 **The control/complexity tension.** The planning and traditional accountability procedures of large organizations are based on control, certainty, predictability, and stability. Complexity resists control, is defined by uncertainty, undermines predictions, and epitomizes turbulence. DE was developed under complexity assumptions. Large organizations operate under control assumptions. These diverse and contrasting orientations create tensions in funding, design, implementation, and reporting.

In summary, developmental evaluation is designed to address issues of emergence or newly arising phenomena and contextual complexity. For the world is changing rapidly, is rapidly becoming more interconnected, and therefore is rapidly becoming more interrelated. These dynamics have many impacts on and consequences for programs, their design, their implementation, their management, their evaluation, and how evaluation can be used to support them.[123]

At CTOP we are particularly interested in how rapid cycle evaluation methods, including those of developmental evaluation, can contribute to building the capacities of programs and the organizations that provide them. Currently we are considering how best to help our grantees adopt such approaches as appropriate to them in building up their programs to improve their quality and effectiveness, and adapt them to the demands of the contexts and systems with which, and within which, they work.[124]

Verstehen versus Erklären – a quick detour through history and back

So far this paper has concerned itself with what, for the most part, are quantitative data. But that isn't, and should never be, the whole story when it comes to evaluation. To get a full picture of what a program is, does, and produces, qualitative data are equally essential.

By the 1970s, the tension between those who insist on quantitative data as the measure of social value and those who see this question as inherently a qualitative matter became intense among evaluators in this country.[125] Whether they knew it or not, they actually were carrying on a fight that had emerged among philosophers in Germany during the late 19th century. It was broached by Johann Gustav Droysen[126] and a bit later by Wilhelm Dilthey,[127] who were intent on legitimating the study of history, philosophy and, more broadly, the humanities against the claims of the natural sciences, which had become the prevailing approach to discovering "truth" and explaining it. Droysen and Dilthey called the aim of natural science "erklären" (which translates from German into English as "to explain"). Against "erklären" they put forth the concept of "verstehen" ("to understand"); this requires a deeper look at things and the meanings human beings place on them.[128, 129] In this dichotomy, which subsequently was brought into the social sciences by Max Weber[130], those people interested in "erklären" focus on quantitative data, those interested in "verstehen" on qualitative data.

The so-called "qual-quant" debate in American evaluation grew to be intense, but has simmered down since the 1970s with a broad acceptance that both qualitative and quantitative data are essential for knowing what's going on in the world, and in particular for knowing about the value of social programs' contributions to society.[131] As the saying goes, **no numbers without stories – and no stories without numbers.** Generally speaking, program evaluations, even those assessing impact, now will used "mixed methods" employing both quantitative measures and qualitative narratives and descriptions. Why? Because numbers without stories are empty, devoid of meaning and human interest. And stories without numbers are just that: stories. They provide no sense of to what degree they are isolated events or illustrate generally prevailing patterns.

---

[123] Baylor, R., Esper, H. Fatehi, Y. de Garcia, D. Griswold, S. Herrington, R., Belhoussain, M. O., Plotkin, G. & Yamron, D. (2019). Implementing Developmental Evaluation: A Practical Guide for Evaluators and Administrators. U.S. Agency for International Development.
Implementing Developmental Evaluation: A Practical Guide for Evaluators and Administrators. (2019) U.S. Agency for International Development.
[124] Gamble, J. A. A. (2008). A Developmental Evaluation Primer. The J. W. McConnell Family Foundation. This is a very useful introduction to developmental evaluation.
[125] Patton, M. Q. (2008). Utilization-Focused Evaluation (4th edition). Sage. See Chapter 12: The Paradigms Debate and a Utilization-Focused Synthesis. pp. 419-469.
[126] Droysen, J. G. (1867). Grundriss der Historik. Veit.
[127] Dilthey, G. (1883). Einleitung in die Geisteswissenschaften. Dunker & Humblot.
[128] Apel, K-O. The Erklären-Verstehen controversy in the philosophy of the natural and human sciences. In: Fløistad G. (ed.). La Philosophie Contemporaine / Contemporary philosophy. International Institute of Philosophy / Institut International de Philosophie, vol 2. Springer. pp. 19-49.
[129] These views were well aligned with "Romanticism", the dominant movement in literature, theater, music, and visual arts in Germany at that time; it was concerned primarily with the emotional or subjective experiences people have and is perhaps best exemplified for American readers by the works of Johann Wolfgang von Goethe (for example, Die Leiden des Jungen Werthers [The Sorrows of Young Werther] and Faust).
[130] See, e.g.: Weber, M. (2002) [1905]. The Protestant Ethic and The Spirit of Capitalism, translated by S. Kalberg. Roxbury.
[131] Patton, M. Q. (2011). Developmental Evaluation: Applying complexity concepts to enhance innovation and use. Guilford Press.

We're Here To Help

CTOP does not expect its grantees to become evaluation experts. But we do expect that they will recognize the importance of using evaluative methods to improve the quality and effectiveness of their work. While CTOP does not have evaluators on its staff (at least at this time), we do have strong relationships with evaluators like Gordon Berlin and Michael Quinn Patton and through such contacts can provide evaluation experts to consult to and support the work of our grantees on an individualized basis. We expect that, as we learn more about rapid cycle evaluation, we will be able collaborate with our grantees in applying evaluative thought and methods to strengthen their programs incrementally until they are ready to undertake more demanding implementation, benchmarking, and (in some cases) impact evaluations.

And we expect that CTOP, too, will become a better social investor and source of support to our grantees as we refine our own evaluative thinking and implement some of the evaluative practices described in this paper.

Connecticut
Opportunity
Project

Learn more at
www.ctopportunityproject.org