A FUNDER'S GUIDE TO USING EVIDENCE OF PROGRAM EFFECTIVENESS IN SCALE-UP DECISIONS

Michael Bangser

May 2014

Social Impact EXCHANGE
TAKING SUCCESSFUL INNOVATION TO SCALE

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

A FUNDER'S GUIDE TO USING EVIDENCE OF PROGRAM EFFECTIVENESS IN SCALE-UP DECISIONS

Michael Bangser

May 2014

Social Impact EXCHANGE
TAKING SUCCESSFUL INNOVATION TO SCALE

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

Funding for this Guide was provided by the Social Impact Exchange at Growth Philanthropy Network and The Annie E. Casey Foundation.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, The Harry and Jeanette Weinberg Foundation, Inc., The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Meyers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

For information about MDRC and copies of our publications, see www.mdrc.org. For information about the Social Impact Exchange at Growth Philanthropy Network, see www.socialimpactexchange.org.

TABLE OF CONTENTS

LIST OF EXHIBITS

Table

Figure

Box

ACKNOWLEDGMENTS

This Guide was written in consultation with the Social Impact Exchange at Growth Philanthropy Network and benefited from the insights of Cynthia Massarsky and Alex Rossides. In addition, helpful comments were provided by many other individuals: Robert Ivry, John Martinez, Howard Bloom, and Michael Weiss at MDRC, as well as Albert Chung, Marcia Egbert, Robert Granger, David Medina, Edward Pauly, Nathan Render, Edward Skloot, and Marta Urquilla.

HOW TO USE THIS GUIDE

This Guide provides funders with practical advice on how to think about and use evidence of effectiveness when considering investments in scale-up opportunities. The Guide does not seek to turn private funders into evaluation experts or to delve into the methodological details of particular research approaches. Rather, the focus is on the right questions that funders should ask and the pitfalls they should avoid, including how to recognize the limitations of certain kinds of evidence. The Guide is divided into three sections:

Section I, Eight Key Questions to Ask Throughout the Scale-Up Process, presents what funders should look for to determine whether programs are effective. These questions provide the building blocks for the discussion in the following section.

Section II, Application of the Eight Questions to Scale-Up Decisions, shows how the questions apply to the different stages of a program's evidence-building and scale-up.

Section III, Next Steps for the Field, highlights some remaining challenges for the field to consider in using evidence of effectiveness to guide scale-up decisions.

The stages of scale-up used in this Guide (early → developing; developing → promising; promising → effective; effective → scaling) are depicted in Tables 1 through 4 (on pages 17 to 20) and in the Appendix. The accompanying text includes suggestions for (1) what the focus of evaluation efforts should be at each stage; (2) how funders can help; and (3) what's needed for the program to move to the next stage of growth. This should help funders to integrate evidence-building into their strategic grant-making process, while recognizing that other factors (such as a grantee's business planning and ability to raise capital) will also influence the prospects for effective scale-up. There are times when scale-up can proceed more quickly and might not require the same level of evidence described in this Guide; however, funders should carefully consider the risks and uncertainties associated with making decisions with more limited evidence.

INTRODUCTION

Foundations and other private funders increasingly seek opportunities to "scale impact" — that is, to extend the benefits of cost-effective interventions to more people, either by expanding these efforts in their current locations or by replicating them in new locations. However, the effect that funders seek to have on vexing social problems, such as entrenched poverty, the educational achievement gap, and health disparities, will not materialize unless they can identify interventions that truly work and then support sustained, high-quality implementation of these interventions as they scale up.

There should be a high bar of reliable evidence to justify substantial scale-up because the stakes are high: Many lives will be affected, substantial funding is typically involved, and valuable resources can be wasted if funders back the wrong interventions. Nevertheless, funders often act without sufficient evidence to guide decisions on whether to invest in particular scale-up opportunities, as well as to confirm that the scale-ups that they do support have been successful.

This Guide focuses on eight key questions that funders should generally ask during the stages of scale-up to help them direct resources to the right places. It includes references to the extensive, thoughtful work done by others,[1] while providing context and recommendations in a format that facilitates funders' use of this material. The Guide draws primarily on lessons and principles from evaluations of *programs* to improve educational, employment, health, and other outcomes for individuals, while recognizing that many funders are also interested in scaling other approaches, including, for example, supporting institutions, disseminating best practices that are embedded in programs with multiple elements, and advocating for changes in public policies and systems.[2] (The lessons in this Guide are relevant to this broad range of interventions, but the terms "program" or "services" are used throughout for the sake of simplicity.)

A key assumption underlies this Guide: Although individual program evaluations need not be unduly expensive or time-consuming,[3] building the evidence to support scale-up decisions is typically not a one-time event. It is a continuous — and typically multi-step — journey involving a variety of partners: those who help produce the evidence (for example, grantees, third-party evaluators,[4] and the agencies that provide relevant data); key public and private decision-makers to whom the evidence must be communicated; and the service providers that scale up evidence-based approaches in a policy environment that is influenced by a range of public officials, advocacy groups, and other constituencies.[5]

Ideally, an effective evidence-building process will:

- Tap the potential for collaborative funding among philanthropic organizations, government, and businesses to support ongoing innovation, evidence-building, and phased scale-up of interventions. To achieve this, funders need to be willing to invest in data collection, analysis, and communication of findings.

1 See the Social Impact Exchange Knowledge Center Web site (www.socialimpactexchange.org/exchange/knowledge-center); Weiss et al. (2010); Major (2011).
2 See Coffman (2010).
3 Baron (2012a).
4 For issues to consider in using third-party evaluators, see Rutnik and Campbell (2002).
5 Fixsen, Blase, Horner, and Sugai (2009).

- Allocate greater resources to organizations that are further along in producing reliable evidence of positive, cost-effective results.
- Promote the use of evidence and quality data collection and monitoring systems to inform mid-course corrections, continuous operational improvement, and sustained impacts as the scale-up of specific programs progresses.

Of course, the complexities of real life mean that scale-up will not necessarily proceed in a predictable manner with all the evidence that might be desirable at each stage. Although funders will need to make judgment calls despite some level of uncertainty, this Guide helps frame the questions and standards that funders should consider in making these judgments.

There are recent examples of how this process can work, with staged scale-up and decision points along the way to guide investment in operations and evidence-building. In the public sector, which often provides critical scale-up funding, the federal Office of Management and Budget is increasingly insisting on rigorous evidence when allocating resources in a tight fiscal environment. Key federal initiatives such as the Social Innovation Fund and the Investing in Innovation (i3) Fund enlist private partners to match public funding in an ongoing process of evidence-building.[6] Philanthropy-led efforts, such as the Edna McConnell Clark Foundation's investment in youth-serving organizations,[7] calibrate funding to the level of evidence while also investing in bringing the evidence to higher levels. The Social Impact Exchange's Funder Working Groups and Scaling Marketplace conduct due-diligence reviews of evidence to support scale-up. The work of Results for America[8] and the Coalition for Evidence-Based Policy,[9] among others, has also highlighted the importance of using reliable evidence to drive policy and practice.

6 See the Corporation for National and Community Service Social Innovation Fund Web site (www.cns.gov/programs/social-innovation-fund); see the U.S. Department of Education Web site (www.2ed.gov/programs/innovation.index.html).

7 See the Edna McConnell Clark Foundation Web site (http://www.emcf.org).

8 See America Achieves Web site (www.americaachieves.org/tools-policy).

9 See the Coalition for Evidence-Based Policy Web site (http://coalition4evidence.org).

SECTION I
EIGHT KEY QUESTIONS TO ASK THROUGHOUT THE SCALE-UP PROCESS

Funders should typically judge the evidence for scale-up by asking eight key questions (listed in Box 1) that go to the heart of understanding a program's effectiveness and readiness for scale-up. As discussed more fully in Section II, these questions are relevant *throughout all stages of scale-up*; however, as scale-up progresses, some of these questions will take on more importance and are expected to be answered with greater rigor.

BOX 1. EIGHT QUESTIONS TO ASK THROUGHOUT THE SCALE-UP PROCESS

1. **What can be learned from existing research and the local context?**
   At the outset, a careful review and synthesis of local conditions and relevant lessons from evaluations of similar programs should feed into development of a sound theory of change.

2. **What are the characteristics of the people being served?**
   A clear understanding of their characteristics and how they were selected helps in interpreting results. A "funnel analysis" can provide important insights.

3. **What is the content of the program, and how well is it implemented?**
   It is important to understand how the theory of change is implemented in practice.

4. **How much does the program improve key outcomes compared with a reliable counterfactual?**
   A reliable counterfactual (that is, a benchmark of how participants would have fared in the absence of the program) is crucial to understanding a program's true value added.

5. **What factors most influence the results?**
   The program elements that contributed most to the positive results should be preserved during scale-up.

6. **How much does the program cost?**
   Consider program costs from a number of perspectives and how the costs might change during scale-up.

7. **How relevant is the evidence to the funder's specific scale-up decisions?**
   Consider how much the accumulated evidence relates to programs operating under the conditions that will be in effect during scale-up.

8. **How can reliable evidence be produced at reasonable cost?**
   Pay attention to the quality of data and how the data are analyzed. Recognize the need to support the real costs of quality evaluations.

Question 1:
What Can Be Learned from Existing Research and the Local Context?

The design and implementation of programs do not begin with a blank slate. There is typically an accumulated body of research on the needs of the population being served, on relevant policy issues, and on operating lessons from similar, or at least related, efforts.

Similarly, programs and related evidence-building are not conducted in a vacuum; their potential for effective scale-up is influenced by the context in which they operate. For example, it is difficult to implement, and especially to scale up, programs that are run by weak or underfunded organizations. Effective scale-up also typically requires strong connections with mainstream funding sources, public policies that facilitate smooth operation of the programs, and relationships that help with client referrals, linkages with complementary services, and other matters. Finally, the organizations, systems of services, and public policies within which programs operate can all influence the ability to produce reliable data on program effectiveness.

Particularly when interventions are replicated in new locations, it is crucial to assess how a different context — the local economy, system of services, and public policies — might affect operations.

Unfortunately, reviews of these issues, sometimes referred to as "scans," typically suffer from two weaknesses. First, they are often not comprehensive enough and lack hard-nosed assessments of the available evidence. Funders should probe the basis for conclusions that a program is, for example, "promising" and ask whether the program is grounded in a credible theory of change — that is, a well thought-out explanation of exactly how a program is supposed to accomplish its goals. Second, the reviews are often conducted later than they should be. For example, the issue of how well a program fits within its organizational structure, the relevant systems of services, and the larger policy environment should be considered from the outset and then continuously monitored as the program evolves. Questions that should be addressed include the following: How likely is it that there will be sufficient, long-term funding for the program if it is scaled? Is the demand for services — along with appropriate outreach and referral mechanisms to attract large numbers of participants — likely to materialize? What organizational capacity, staff development, and data systems will be needed to support scale-up? Do applicable government regulations promote or constrain effective program operations? Particularly when interventions are replicated in new locations, it is crucial to assess how a different context — the local economy, system of services, and public policies — might affect operations.

The results of the review should feed directly into the development of a sound theory of change that defines the specific outcomes being sought and provides a compelling basis for the pathways to achieve those outcomes. In addition to guiding program operations, a clear theory of change helps to focus data collection and analysis, including identification of possible reasons for the program's success or failure.

Key elements to look for are:

1. Specification of the program's desired outcomes.
2. Clearly articulated assumptions about the characteristics of those being served and the context in which the program operates.
3. A framework that connects the resources and program elements needed to achieve the outcomes.

4. Indicators to measure the intensity and duration of individuals' program participation and the outcomes they achieve.

5. A narrative or graphic presentation that explains the logic and sequencing of the program.[10]

As discussed in Section II of this Guide, the theory of change may — and often should — evolve, especially as more is learned in the early stages of program implementation.

Question 2:
What Are the Characteristics of the People Being Served?

Information on the characteristics of the individuals that a program serves is critical for two principal reasons:

- To confirm that the theory of change, the basic design of the program, and its operating structure (for instance, staffing patterns) are appropriately tailored to the needs of the intended beneficiaries.

- To properly interpret data on program operations and outcomes, since ostensibly favorable (or unfavorable) results could reflect the characteristics of the people being served as much as, or perhaps even more than, how well the services are provided. Particularly relevant here are those characteristics of individuals that are likely to be correlated with the intended outcomes. For example, it might be easier for a job training program to place participants with strong employment histories, education levels, and motivation to work than those without such qualities. (See Question 4 for more on this issue.)

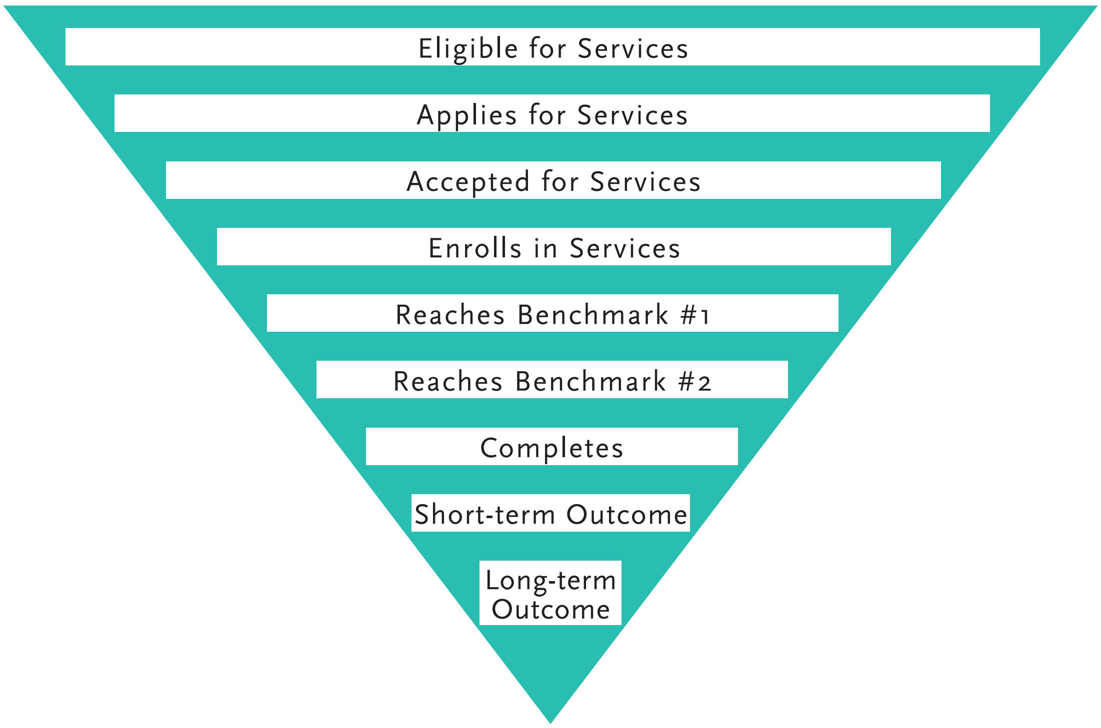
It is also important to understand *how participants were selected* into the program and the extent to which the selection process may have screened out harder-to-serve individuals — thereby admitting only those most likely to succeed whether they participated in the program or not. A "funnel analysis" can help clarify this point.

As depicted in Figure 1 (page 6), a funnel analysis uses reliable data collected on the universe of individuals who satisfy program eligibility requirements, on key steps in the recruitment and selection process (such as interviews or testing), on the extent to which particular types of individuals progress through the various steps, and on the reasons for any drop-off along the way, especially whether the program staff exercised discretion to exclude certain people who were considered more difficult to serve. (A funnel analysis will also be useful to understand participation levels and drop-off points *during* the program. For example, in a program with a sequence of components, did participants make the transition from one component to the next? Did they participate for the number of hours of program activity that the theory of change assumes is needed to achieve the desired outcomes?)

It should be noted that the characteristics of individuals served may well change over the course of scale-up, either because replication brings the intervention to new communities or because expansion in the current location means that there will be a need for outreach to a broader (and sometimes more difficult to serve) range of individuals.

10 See the Center for Theory of Change Web site (www.theoryofchange.org); Reisman, Gienapp, Langley, and Stachowiak (2004); W. K. Kellogg Foundation (2004).

FIGURE 1. ENROLLMENT AND PARTICIPATION FUNNEL



Question 3: What Is the Content of the Program, and How Well Is It Implemented?

To determine how well the theory of change is being executed in practice, funders will want to understand how the services are *actually provided*, as well as whether there are any changes over the course of scale-up. In particular, funders should focus on program operating data and implementation research that address the following:[11]

- *Program content*: What specific curricula, techniques, incentives, and other approaches are used? Is the content of the program properly aligned with the specific needs of the population being served?
- *Quantity or "dosage" of services provided:* How often (frequency), for how long (intensity), and over what period of time (duration) are services offered and received? In particular, does the funnel analysis described earlier confirm that individuals are *actually engaged* at the levels that are likely to be needed to produce the desired outcomes? If, as is often the case, there is a shortfall in this threshold performance measure, substantial investment in scale-up will not be appropriate unless corrective action leads to measurable improvement.
- *Quality*: This is more difficult to measure than quantity, but certain factors thought to affect quality can be quantified (for instance, teacher:child ratios in preschool programs), and

11 This discussion draws heavily on Weiss, Bloom, and Brock (2013) and on Durlak and DuPre (2008).

there are instruments to assess quality in many areas such as preschool,[12] after-school,[13] summer learning,[14] and home-visiting programs.[15]

- *Means of delivery*: What methods (for instance, in-person or by phone, individually or in groups) are used to provide the services? Note that the means of delivery can influence the quantity, quality, and cost of a service.

Each of these elements could be influenced by the pressures arising from scale-up. For example, there could be challenges in tailoring services to a different client mix (whether in a new location or because of outreach to previously unserved groups), in recruiting staff, and in garnering full funding for all the program components.

Question 4: How Much Does the Program Improve Key Outcomes Compared with a Reliable Counterfactual?

Funders naturally want to know whether particular services make a difference on measurable outcomes of interest. The best way to determine that is to develop a reliable "counterfactual" — that is, a measure of what would have happened in the absence of the services. A particular challenge is that confirming a *reliable* counterfactual can be especially difficult for certain types of interventions, such as those seeking community-level impacts. In the absence of a reliable counterfactual, funders should acknowledge the level of uncertainty — and potential for downright misleading results — that remain about a program's effectiveness. The remainder of this section discusses three issues to consider: (1) distinguishing between gross outcomes and net impacts; (2) taking care in interpreting results; and (3) assessing the magnitude of net impacts.

Distinguishing between gross outcomes and net impacts

To achieve meaningful change at scale, an effective program should reach people who are most likely to *benefit from the services*. It is crucial to define "benefit" as the value added that the program actually caused — that is, the difference it made over and above how the individuals would have fared without being offered the program. To determine this value added, *gross outcomes* (such as how many trainees in an employment program obtain jobs paying at least $15 per hour) must be distinguished from *net impacts* (the extent to which these outcomes were actually caused by the intervention being funded). For example, knowing that 80 percent of participants entered employment at the target wage is meaningful only if there is a reliable measure (the counterfactual) of how many of them would have done so anyway.

Rigorous evaluations using random assignment research designs[16] are widely considered to be the most reliable way to identify an accurate counterfactual, to measure net impacts, and to avoid a

12 Pianta et al. (2008).

13 For instance, The Youth Program Quality Assessment: A Research-Validated Instrument and Comprehensive System for Accountability, Evaluation, Program Improvement. See High/Scope Educational Research Foundation (2012) and the High/Scope Educational Research Foundation Web site (www.highscope.org).

14 See the National Summer Learning Association's Comprehensive Assessment of Summer Programs Web site (www.summerlearningassociation.org).

15 See the Home Visiting Evidence of Effectiveness Web site (http://homvee.acf.hhs.gov); Duggan et al. (2007).

16 Random assignment is essentially a coin flip. For programs that do not have the capacity to serve everyone who is interested, applicants are assigned to either a group that receives the program being tested or to a control group that does not take part. (The control group is typically free to participate in other services available in the community.) If the assignment process is truly random and the research sample is large enough, there should be no systematic pre-program differences between the participant and control groups. Any statistically significant differences in outcomes can therefore be confidently attributed to the effects of the program.

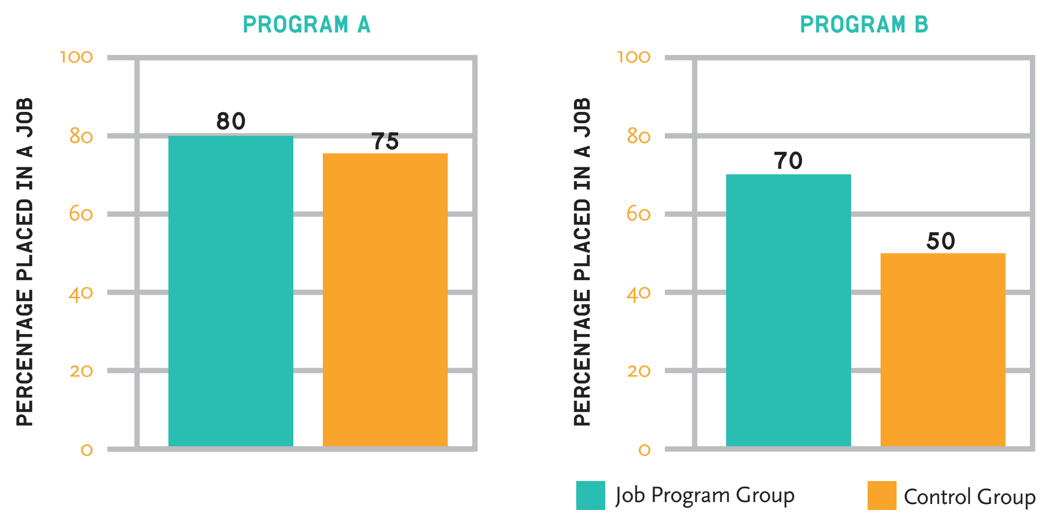
potentially serious misallocation of resources.[17] Figure 2 illustrates the importance of distinguishing gross outcomes from net impacts. It shows, for two different job training programs, the percentage of program participants and the percentage of a randomly assigned control group who began employment during the period of the evaluation. While 80 percent of the participants in Program A found employment, this percentage is only modestly higher than the 75 percent who would have found jobs without the program services, as measured by the experience of the control group. In Program B, 70 percent of somewhat harder-to-serve participants found a job — 20 percentage points higher than in the control group. Thus, the gross outcome was clearly higher in Program A, but the net impact — what funders should care most about — was greater in Program B.

While this pattern does not always hold true, it occurs frequently enough to underscore a crucial cautionary note: a sometimes surprisingly high percentage of program participants (as measured by the experience of the randomly assigned control group) would have made at least some progress even without entering the program being evaluated. For example, a rigorous evaluation of Career Academies — a popular high school reform that combines academics with career-development opportunities — found that while more than 90 percent of the program participants graduated from high school, an equal number of students in the control group also did so. Career Academies did, however, produce substantial and sustained net impacts on the post-high-school earnings of the students in the program.[18]

Better-than-expected outcomes for a control or comparison group may frequently result when a program attracts only the most qualified or motivated individuals, or when a service provider actively screens out harder-to-serve individuals. One way to surface the level of screening is to conduct the kind of funnel analysis described earlier in the discussion under Question 2.

On a related point, the value added of a program may well be greater for more at-risk individuals, who have less favorable gross outcomes but for whom the net impacts could nevertheless be substantial. For example, in a random assignment study of Upward Bound, participants who initially had low

FIGURE 2. ILLUSTRATION OF JOB PLACEMENTS



17 Gueron (2005).
18 Kemple (2008).

expectations of obtaining a bachelor's degree were indeed less likely to go on to four-year colleges than those with higher initial expectations (38 percent compared with 56 percent). However, the Upward Bound program made no significant *difference* for the students with higher expectations, while it more than *doubled* the rate for students who started with lower expectations.[19]

Taking care in interpreting results

Other approaches besides random assignment have been used to identify the counterfactual;[20] however, most data available to funders on the effects of particular services either have no counterfactual or one that uses a comparison group of individuals who differ in important respects from the individuals served. An inappropriate comparison group can lead to two common problems that produce misleading conclusions:

- *Selection bias*. This occurs if, at the point of program entry, the individuals in the program group differ systematically from those in a comparison group on factors that could influence the outcomes of interest. Selection bias is especially likely when participation in services is voluntary, since the process by which individuals learn about, volunteer for, are accepted by, and actually enroll in services (essentially what was reflected in the first part of the funnel analysis described earlier) is exceedingly difficult to match in selecting the comparison group. The program and comparison groups may well differ in important respects, such as motivation levels, that are difficult to measure. This can occur, for example, in studies of substance abuse programs, charter schools, or other situations in which the motivation of participants or the strength of pre-existing support networks can influence later outcomes. Without a reliable counterfactual, the potentially significant effects of these factors will not be accounted for properly.

- *Maturation or other natural changes over time.* In some cases, the observed changes in outcomes would have occurred during the relevant follow-up period even without a program intervention.[21] For example, young children improve on developmental measures as they age, and employment outcomes for adults could especially improve in a rising economy. In cases such as these, pre-post measures can lead to a misleading conclusion that the special intervention, rather than natural maturation or other developments, caused the positive changes.

As noted earlier, in the absence of random assignment or another acceptable approach to identify the counterfactual,[22] funders should acknowledge the considerable uncertainty that remains about a program's effectiveness. If funders have only outcome (rather than net impact) data, they should at least seek to understand (1) the conditions in which the services are provided (for instance, how the local economy might affect an employment program); and (2) the characteristics of the individuals served and how they were selected for the services. Both of these factors provide a context for interpreting gross outcome measures.

19 Seftor, Mamun, and Schirm (2009).

20 It is now generally thought that the next best thing to a randomized trial for estimating program impacts is a regression discontinuity design. For a discussion of this quasi-experimental alternative, see Bloom (2012) or Imbens and Lemieux (2008).

21 Of course, sometimes the outcomes for participants would have gotten worse, not better, in the absence of an intervention. A classic example is "summer learning loss," in which students, especially those from low-income families, score better on reading and math tests at the end of one school year than they do at the beginning of the next school year.

22 See Baron (2012b) for a checklist on what to look for to judge the reliability of a quasi-experimental design.

Assessing the magnitude of net impacts

Funders will make their own judgments on whether the net impacts of particular services are positive enough to justify investment in further scale-up. This assessment should, however, consider:[23]

- *The extent to which the services achieve a specific policy goal.* Such an objective might be to close the well-documented achievement gap between low-income students and more affluent students.

- *How the net impacts compare with the natural growth or change that would normally be expected for the relevant population*. For example, how much does an educational intervention accelerate the learning gains that would normally occur for a given group of students during one school year?

- *How the impacts compare with the effects found in past evaluations of programs for similar populations*. However, in judging effectiveness relative to results for other efforts in the field, funders should ensure that there is an "apples to apples" comparison, taking into account the lessons from this Guide — in particular the distinction between gross outcomes and net impacts.

- *How the net impacts relate to the cost of the intervention that produced them*. This subject is addressed in the discussion under Question 6.

In addition, funders need to understand the exact nature of the counterfactual, particularly the extent of the "service contrast"— that is, how much more and how different are the services in the program that is being tested relative to the services received by members of the control or comparison group? This point is especially important because most real-world evaluations do not test a program against a "no treatment" control or comparison group. Especially if the control or comparison group receives substantial services or ones that are similar to those being tested, the net impacts of the program will likely be dampened. The evaluation would, in reality, be testing the *incremental effect* of the program versus the mix of services received by the control or comparison group. Evaluating the incremental effect could still be relevant — if, for example, the question of interest is whether a new approach is better than the status quo — but the key is for funders to have a clear understanding of what the evaluation is actually testing.

Question 5:
What Factors Most Influence the Results?

Most evaluations focus on the effectiveness of the overall package of services but have difficulty isolating the contribution made by particular elements, such as the separate parts of a multi-component program or the effect of different staffing patterns. Funders will therefore typically have the most confidence in acting on evidence for the program as a whole.

However, there is often insufficient funding to scale up the full program model. For this and other reasons, funders considering scale-up investments will want to help preserve the elements of the intervention that contributed most to the positive results and to discard, replace, or improve the

23 The first three approaches are drawn from Bloom, Hill, Black, and Lipsey (2008).

elements that limited effectiveness. Important methodological advances have been made in ways to identify these elements,[24] but identifying the factors that drove the results is still often more art than science.

Funders who are grappling with this issue should, in consultation with their evaluation partners, consider:

- Focusing on the theory of change, which describes the expected pathways to achieving positive outcomes and can provide clues about possible explanations for the results. For example, were the key elements delivered in the manner that was envisioned, in both quantity and quality?

- Returning to the funnel analysis, described earlier, which might help to identify bottlenecks or drop-off points in the flow of participants into and through the services. If, for example, there were low participation rates in certain program components, it is unlikely that they contributed significantly to any positive results.

- Using interviews, focus groups, or surveys to learn the perspectives of the staff who delivered the program and of the individuals who were served.

- Looking for variations in how and where a multi-site program was implemented as sources of possible explanations (that is, attempting to learn as much as possible from natural program variation).

- Drawing on the lessons learned from evaluations of similar programs serving the same population. (As noted at the outset of this Guide, evidence-building is an ongoing process that benefits from the cumulative insights developed over time in a particular field.)

- Thinking about the possibility of rigorously comparing the impacts of alternative program variations through random assignment of clients to these alternatives in instances where the stakes are high enough.

Most evaluations focus on the effectiveness of the overall package of services but have difficulty isolating the contributions made by particular elements.

Question 6: How Much Does the Program Cost?

Of course, cost is of particular importance to funders, who will be footing the bill for scale-up. Consideration of program costs will be relevant from the inception of a program, but more comprehensive analysis is expected as the program model solidifies and as scale-up progresses. In addition to understanding the costs of the intervention for budgeting purposes and to compare these with the expense of other programs, funders might be especially interested in two other measures:

- **Cost-effectiveness analysis**:[25] This is the cost per unit of outcome (for example, per placement of a participant in a job or per high school graduate). It provides a useful benchmark for comparisons with alternative approaches serving the same population.

- **Benefit-cost analysis**:[26] With appropriate adjustments for discounting of future benefits and for inflation, this provides a basis to compare the upfront investment in a program intervention with the benefits that are spread over time. Since benefit-cost analyses are

24 Bloom (2005).
25 Siegel, Weinstein, Russell, and Gold (1996).
26 See Cellini and Kee (2010).

expressed in monetary terms, there are special challenges for impacts on certain quality-of-life or other measures that are not readily monetized; these items will therefore need to be assessed qualitatively, outside the formal benefit-cost calculation.

Particular points for funders to consider include:

- Since the costs of a program are typically incurred up front, but the benefits can accrue over a multi-year period, funders will often need to determine their comfort level with extrapolations of longer-term benefits from short-term data.
- A benefit-cost analysis or any other analysis of return on investment depends in significant part on assumptions regarding the net impact of the program. Funders should especially consider the need for a reliable counterfactual and the potential for misleading interpretation of gross outcomes rather than net impacts. (See the discussion under Question 4.)
- In some cases, a significant component of the benefits is the reduced use of publicly funded services, such as special education or prisons. A relatively expensive program could pay for itself if it leads to reduced future use of these costly services by especially disadvantaged populations. Alternatively, programs that have relatively modest net impacts might produce a positive return on investment if the cost of the program is low.
- The cost structure of programs will often change during different stages of scale-up. Unit costs tend to be high during the start-up phase (whether in the original location or when replicated) and to decline as the program moves to a larger scale, when capital and overhead costs can be spread over larger numbers of participants and economies of scale can potentially be realized.
- It is often difficult to sustain full program funding levels over an extended period of time or when replicating programs in new locations. This obstacle could create the need to reduce costs and change how the program is operated. These decisions should be driven in large part by the assessment of the program elements that contribute most to positive results. (See the discussion under Question 5.)

Question 7:
How Relevant Is the Evidence to the Funder's Specific Scale-Up Decisions?

Since most evaluations provide findings on the effectiveness of a program for particular types of individuals in particular locations, funders will want to understand how well evaluations conducted in certain contexts will predict the results for the specific individuals and locations that the funders care most about. To reach this judgment, funders should consider whether:

- The people receiving the services that were evaluated are similar to those who will be served when the program is expanded or replicated, since some programs are more effective for certain types of individuals than for others. Particular attention should be paid to (1) the characteristics that would be correlated with the outcomes of interest, such as employment history and education levels for participants in a job training program; and (2) whether the recruitment and enrollment process, including the level of screening of enrollees, will be the same in the original and in the scaled-up services.

- The locations involved in the evaluations were similar in material respects to those where scale-up would take place. A clear example is the unemployment rate for the locality in which a job training program operates. Particular attention should also be paid to the policy environment and systems in which the services are offered, as well as the level of services received by individuals in a control or comparison group.

- The scale-up process itself might affect the content, quantity, quality, or means of delivering the program. (See the discussion under Question 3.)

- The findings were consistent across a variety of population groups and locations. Consistent findings might suggest that the program is adaptable to diverse conditions, potentially making the results of replication in new locations more predictable.

Question 8:
How Can Reliable Evidence Be Produced at Reasonable Cost?

The reliability of an evaluation depends in large part on the quality of the data used. Collection of quality data on the range of desired outcomes is often the most expensive part of an evaluation effort. In general, as the scale-up process unfolds, expectations for the scope and quality of data will increase, with a commensurate increase in staff time and in the financial investment needed for collection, analysis, reporting, and use of the data.

Funders will rely heavily on grantees and evaluation experts in this area but should keep the following considerations in mind:

Within the inevitable budget constraints, funders should focus on ensuring a core set of reliable data on high-priority operating and outcome measures.

- Early on, funders should determine whether service providers have a reliable management information system or other method to collect, to analyze, and to use basic performance measures (for example, the characteristics of enrollees, the service levels they actually receive, and their outcomes). Reliable data for a funnel analysis (see Question 2) will help identify the degree to which the service provider is screening out certain individuals and the points at which participants are dropping out or failing to receive the intended level of services.

- Poor-quality data can lead to faulty or misleading conclusions. Within the inevitable budget constraints, funders should focus on ensuring a core set of reliable data on high-priority operating and outcome measures that (1) relate to outcomes that the program can realistically influence within the relevant time frame; and (2) will actually influence the service providers' and funders' decision-making.[27]

- Grantee-provided data are typically more reliable when the grantees view the data as useful for their own internal operations. Therefore, funders should encourage the grantees' use of data for continuous quality improvement.

- Evaluations conducted by independent third parties can provide helpful perspective and objectivity; however, the fact of a third-party evaluation does not, by itself, guarantee that the data and analysis are reliable.

27 See Twersky, Nelson, and Ratcliffe (2010).

- Since data collection and reporting requirements can impose burdens on grantees and others on the front lines, funders should consider providing resources to support these activities and joining in efforts to define common measures and reporting procedures acceptable to multiple funders.

- Some outcomes are more easily quantifiable than others; it is not always easy to measure the results that matter most. Funders should use the metrics that are available to narrow the range of uncertainty, to acknowledge the unknowns, and then to apply informed judgment in making final decisions about whether and how to proceed with scale-up.

Funders are understandably concerned about the cost and timeliness of evaluations, some of which can be expensive, but there are ways to manage these costs. Among these are:

Funders are understandably concerned about the cost and timeliness of evaluations, some of which can be expensive, but there are ways to manage these costs.

- As discussed in the next section, the scope of (and level of investment in) evaluation should be calibrated to the stage of scale-up. If, for example, there is encouraging evidence from operational indicators and from an outcome study, it could be time to fund a study of net impacts to confirm whether an intervention *caused* the positive outcomes.

- Lessons can be drawn from other reliable evaluations using similar populations, such as those reported in the What Works Clearinghouse and in Social Programs That Work developed by the Coalition for Evidence-Based Policy. However, funders should be careful to assess how much the evaluation findings from a particular context provide reliable evidence for the funders' decision-making. (See the discussion under Question 7.)

- Discuss the possibility of joining with other funders to support evaluation costs in areas of common interest.

- Consider lower-cost evaluation options that use administrative data, as has been done, for example, in random assignment studies in the education and criminal justice fields.[28] (Evaluation costs are not driven primarily by whether it is a random assignment study, but rather by factors such as the expense of data collection, how many sites are involved, and the length of follow-up for the research sample.) Administrative data probably will not, however, provide information on the full range of outcomes of interest to funders. Administrative data also will not answer important questions about *how* the program was implemented and *why* an outcome was or was not achieved.

28 See Baron (2012a).

SECTION II
APPLICATION OF THE EIGHT QUESTIONS TO SCALE-UP DECISIONS

Scale-up of an intervention should be considered a multi-stage process. Various authors and organizations have proposed somewhat different frameworks to define the number, duration, and nature of these stages,[29] but there is generally a progression from the early conceptualization of a program through various stages of implementation and evaluation, and, if warranted, to larger-scale expansion or replication. For the purposes of this Guide, the stages of evidence-based scale-up are:

Stage 1: Early → Developing
Stage 2: Developing → Promising
Stage 3: Promising → Effective
Stage 4: Effective → Scaling

There are not strict lines of demarcation between the stages, which may overlap, and scale-up obviously does not always proceed in the straightforward manner presented here. Indeed, programs sometimes scale up — either by expanding in a single site or by operating in numerous locations — before there is reliable evidence of program effectiveness. Funders might therefore need to decide early on whether there is sufficient evidence of effectiveness to justify investing in something that has *already* been scaled up to a certain extent. There may be instances in which scale-up is appropriate without having the kind of net impact evidence, with counterfactuals, discussed in this Guide.[30]

In the sequence described below, a primary goal of each stage is to build evidence and to strengthen program operations in order to determine whether it is appropriate to move to the next stage. For many programs, scale-up will never be appropriate, either because of the lack of evidence of effectiveness or because of weak business plans or limited capacity in other respects. The eight evaluation questions discussed in Section I of this Guide (see Box 1) should be kept in mind throughout the scale-up process. As noted earlier, the expected range and rigor of evidence, as well as the level of investment devoted to producing that evidence, normally increases as a program proceeds from one stage to the next. Therefore, at each stage, funders should consider not only the effectiveness of the program but also the capacity to conduct the more rigorous data collection, program monitoring, and evaluation (as well as business planning and organizational development) that is envisioned for the next stage. In practice, the principles set forth in this Guide will be implemented in various ways to take account of real-world conditions. Box 2 (page 16) provides three examples.

The discussion below summarizes the nature of each stage of the scale-up process and identifies (1) what the priority evaluation considerations should be; (2) how funders can help to build the desired evidence; and (3) what is needed to move to the next stage of the scale-up process. To demonstrate the flow from stage to stage, the full sequence of evidence-building is shown in the Appendix. In many cases, the steps identified in one stage will need to be continued into the next stage.

---

29 See, for example, The Edna McConnell Clark Foundation Web site (www.emcf.org/our-strategy/our-selection-process/evidence); McDonald (2010); U.S. Department of Education i3 Web site (www.2.ed.gov/programs).

30 For example, some funders might be willing to assume that the benefits of well-run programs that meet basic human needs, such as food or clothing distribution, are self-evident. Their evaluation questions might then focus on confirming the efficiency and cost-effectiveness of distribution, the absence of duplicative services, and the quality of the food or clothing that is distributed.

BOX 2. THE ROAD TO SCALING UP: THREE PROGRAM EXAMPLES

The following are three examples of the link between evidence-building and scale-up. The first two programs have reached substantial scale, while the third is at an earlier stage.

Nurse-Family Partnership (NFP; www.nursefamilypartnership.org), a maternal and early childhood health program, relies on specially trained public health nurses who regularly visit low-income expectant mothers during their first pregnancy and the first two years of their children's lives. Grounded in self-efficacy, human ecology, and attachment theory, NFP is a carefully developed model that has been tested and refined over time, including through three randomized control trials in varied settings. Replication of the model reflects the evaluation findings (for example, by insisting on use of specially trained nurses with a baccalaureate degree in nursing, even though they are more expensive than paraprofessionals with more limited training), and benefits from extensive support by NFP's National Service Office.

NFP is now operating in more than 40 states, in many cases under the federally funded Maternal, Infant and Child Home Visiting Program, and with general operating support and growth capital provided by a number of private funders. The commitment to evaluation has continued even after scale-up: NFP is now one of four programs participating in a major evaluation that will provide further evidence on how the programs are currently operating (such as how many home visits families receive and how much training and supervision is provided), and use larger and more demographically diverse samples to explore program effects that may not have been seen in prior studies. (For a detailed review of the NFP evaluation evidence, see NFP's Web site and the Web site of the Coalition for Evidence-Based Policy: http://toptierevidence.org/programs-reviewed/interventions-for-children-age-0-6/nurse-family-partnership.)

Multisystemic Therapy (MST) is an intensive family- and community-based treatment program that focuses on the systems that impact chronic and violent juvenile offenders — their homes and families, schools and teachers, neighborhoods, and friends. Evidence of MST's effectiveness in treating chronic juvenile offenders led to interest in exploring whether the techniques would also be effective for other populations, such as individuals with psychiatric or substance abuse problems. These "adaptations" are supported by MST Services, which provides structured training, coaching, and other resources to ensure that the model is implemented with fidelity. The adaptations typically follow a sequence beginning with relatively low-cost pilot studies to determine feasibility and preliminary effects. This is followed by a series of trials under carefully controlled conditions and then in more real-world community-based settings. Broader dissemination (called "mature transport") occurs when MST Services is reasonably confident that the intervention protocols will achieve the desired outcomes if implemented with fidelity, and that the training and quality assurance procedures are sufficient to support the effective implementation. (See http://mstservices.com.)

The **Accelerated Study in Associate Programs** (ASAP) is at an earlier stage of evidence-building. ASAP requires students to attend college full time and provides a comprehensive array of supports and incentives for up to three years to help developmental (remedial) education students in six New York City community colleges earn their degrees in a timely manner. An internal evaluation by the City University of New York showed promising results, prompting interest in an independent, rigorous random assignment evaluation. Early results show that, compared with a control group of students receiving their college's standard services, ASAP students had substantially higher graduation and two-year graduation rates. Since the comprehensive, multi-component ASAP program is more expensive than standard services, the ongoing research agenda includes additional cost analysis that will be important in making decisions about further scale-up of the ASAP model. It will also be important to test the model in community college systems beyond New York City. (For more on the ASAP evaluation, see www.mdrc.org/publication/more-graduates and www.mdrc.org/publication/what-can-multifaceted-program-do-community-college-students.)

TABLE 1. STAGE 1 OF THE SCALE-UP PROCESS
EARLY → DEVELOPING
(PRE–SCALE-UP)

PRIORITY EVALUATION CONSIDERATIONS	HOW FUNDERS CAN HELP	WHAT'S NEEDED TO MOVE TO THE NEXT STAGE
• Understand the needs and opportunities of the population being served and how they are selected *(Questions 1 and 2)* • Conduct a careful review of local conditions, operating lessons, and results of evaluations of similar programs for similar populations *(Question 1)* • Explore alternative approaches, develop a theory of change and program model, and consider early implementation lessons *(Questions 1 and 3)* • Track participation characteristics, early performance measures (especially participation levels in key program components), and outcomes *(Questions 2, 3, and 8)*	• Connect grantees with key constituencies (including the people whom the intervention will serve) and experts in the field to understand the local context and lessons from previous evaluations • Ensure that the information collected is synthesized and used to help sharpen the theory of change and to develop the program model • Encourage operating flexibility and experimentation • Support development of a performance tracking system (with unique participant identifiers) that is tied to the theory of change • Begin to understand potential contributions of specific program components	• A plausible theory of change that is grounded in the available research and in the early operating experience; it should show potential pathways to success, beginning with how participants are recruited and enrolled and extending through the mix and sequence of services that are expected to produce the desired outcomes • A reasonably well-developed participant tracking system • Initial anecdotal and other evidence suggesting that the program is feasible to operate and has genuine potential to produce the desired results

Stage 1: Early → Developing

Stage 1 is a period of exploration that involves understanding the characteristics of the population being served, experimenting with different approaches, developing a plausible theory of change, and developing the design of the program model. (See Table 1.) As described by Preskill and Beer, particularly in truly innovative efforts, this stage could lend itself to what is known as "developmental evaluation," which emphasizes operational flexibility and adaptation to local conditions.[31] At this stage, the program is typically operating in one or in a small number of locations for a limited number of participants and is not ready for scale-up. Indeed, many programs stay at this or the next level (Developing → Promising) and will never be ready for substantial scale-up.

Stage 2: Developing → Promising

In Stage 2, a sound theory of change and a reasonably defined program design should be in place, along with encouraging early operational experience on threshold questions, such as whether the program is engaging its intended beneficiaries with the expected content, quantity, and quality of services. At this stage, the program is ready to be tested, though typically in relatively small-scale demonstration efforts in a single or limited number of sites in order to better understand its potential. As illustrated in Table

31 Preskill and Beer (2012).

TABLE 2. STAGE 2 OF THE SCALE-UP PROCESS
DEVELOPING → PROMISING
(PRE–SCALE-UP)

PRIORITY EVALUATION CONSIDERATIONS	HOW FUNDERS CAN HELP	WHAT'S NEEDED TO MOVE TO THE NEXT STAGE
• A funnel analysis reflecting steps in the recruitment and selection process, the extent to which individuals progress through the steps, and the reasons for any drop-offs, especially whether program staff excluded certain people considered more difficult to serve *(Question 2)* • Further development of the theory of change, based on early operating experience *(Question 3)* • Documentation that threshold levels of participant engagement have been reached • Establishment of specific outcome goals that are reasonable in light of the participant characteristics and selection process; documentation of progress in achieving the outcome • Initial cost estimates, but not yet cost-effectiveness or benefit-cost figures	• Support improvements in the participant tracking system, including data needed for a funnel analysis and continuous program improvement (for example, ways to increase participation in key program components) • Possibly fund independent evaluator for formative, implementation, or outcome analysis; net impact studies typically not conducted yet	• A better understanding, based on early operating experience, of the organizational, systems, and policy context for the program • Understanding of participant characteristics (for program design and interpreting outcomes) • A reasonably well-developed theory of change • A functioning internal system to capture participant characteristics and to track participant activity and outcomes • Positive internally generated operating and outcome data demonstrating that participation levels and outcomes are consistent with a theory of change • Acceptable operating costs

2, evaluations at this stage should normally use a reasonably well-developed internal management information system to document key operating indicators such as participant activity levels, basic cost estimates, and outcome measures. An independent, third-party evaluator might be engaged for "formative,"[32] implementation, or outcome analyses (or any combination of these). Depending on the results, the program could be ready to go to the next stage, possibly with some adjustments to the theory of change or to the program model.

Stage 3: Promising → Effective

In Stage 3, an independent evaluation should determine more rigorously whether the program *caused* the observed outcomes in real-world conditions and perhaps on a modestly expanded scale, as described in Table 3. For example, an early childhood or after-school program might be tested more broadly within the regular service-delivery system with typical providers, rather than in a special demonstration context operated only by "early adopters." The analysis of a program's net impacts

32 A formative evaluation focuses on early diagnosis and feedback to improve program operations.

TABLE 3. STAGE 3 OF THE SCALE-UP PROCESS PROMISING → EFFECTIVE

(LIMITED SCALE-UP/REPLICATION MAY BE APPROPRIATE AS PART OF LEARNING AGENDA)

PRIORITY EVALUATION CONSIDERATIONS	HOW FUNDERS CAN HELP	WHAT'S NEEDED TO MOVE TO THE NEXT STAGE
• Testing operations more broadly in the regular service delivery system • The extent to which the program *causes* improved outcomes, using random assignment or strong quasi-experimental design *(Question 4)* • Complementary implementation and cost analyses, perhaps including cost-effectiveness or benefit-cost studies *(Questions 3, 5, and 6)* • Assessment of fidelity to the key elements of the model, although model refinement continues *(Questions 3, 4, 5, and 7)*	• Support impact, implementation, and cost evaluations by independent, third-party evaluator(s) • Support analysis of which components should be implemented with fidelity and which are better left to local adaptation • Support dissemination of evaluation findings • Convene potential funding partners • Provide capacity-building grants to prepare for scale-up • Provide capacity-building grants to strengthen data collection and analysis for continuous quality improvement	• Positive net impact results *(Question 4)* • Sound theory of change and clear articulation of key program elements • Strong internal system to track participant characteristics and activity levels, staff activity, and program outcomes in context of program expansion and/or replication • Reasonable fidelity across program sites and quality-control measures to maintain fidelity • Strong support for scale-up within the operating organization(s), system of services, and policy/funding environment *(Question 1)*

(using a random assignment or a strong quasi-experimental design) should be complemented by implementation and cost analyses, although it may be premature to conduct a full-scale benefit-cost analysis. Typically, at this stage the basic program design is reasonably settled and fidelity (or adherence) to the design is examined. However, there is still room to refine the design. In addition, there needs to be an appropriate balance between fidelity to the original program model and adaptation to the changing local conditions or populations served in scale-up. (Since, at this stage, it may be that the program has already been scaled to some extent, care should be taken to confirm that the scaling has not outpaced the level of supporting evidence. If scaling has indeed occurred, portions of the Effective → Scaling section below become relevant.)

Stage 4: Effective → Scaling

During Stage 4, the intervention is scaling, and the evidence-building agenda seeks to confirm effectiveness in multiple settings and for larger numbers of individuals, as indicated in Table 4 (page 20). The scope and rigor of evidence, as well as the corresponding investment in evidence-building, should be calibrated to the degree of scale-up and to the level of investment, but random assignment or strong quasi-experimental designs should be included. Funders should pay particular attention to how scale-up might influence the characteristics of participants, as well as the capacity of program operators to maintain or increase net impacts when the program reaches substantial scale in a single location or in multiple locations. The issue of fidelity is examined closely to determine which elements of the intervention should be held constant and which should be tailored to varied local circumstances. There is a corresponding need for effective systems to promote quality control and consistency of impacts across multiple sites.

TABLE 4. STAGE 4 OF THE SCALE-UP PROCESS
EFFECTIVE → SCALING
(SCALE-UP RANGING FROM MODEST TO EXTENSIVE IN NUMBER OF LOCATIONS AND SIZE OF OPERATIONS IN EACH LOCATION)

PRIORITY EVALUATION CONSIDERATIONS	HOW FUNDERS CAN HELP	WHAT'S NEEDED TO MOVE TO THE NEXT STAGE
• The effect of scale-up on participant characteristics *(Question 2)* • Impact and implementation analyses, focusing on maintaining or improving results in context of scale-up, potentially across numerous locations *(Questions 3, 4, 5, and 7)* • Cost-effectiveness and/or benefit-cost analysis *(Question 6)* • Refined analysis of appropriate balance between fidelity and local variation • Plans for continued evaluations	• Support impact, implementation, and cost evaluations by independent, third-party evaluators; include cost-effectiveness and/or benefit-cost studies • Support development of a "fidelity guide" clarifying elements to control and elements for which variation is appropriate • Provide capacity-building grants to sustain scale-up • Provide continued support for dissemination of evaluation findings • Continue to help convene funding partners	• Calibration of the rigor of evaluation and corresponding investment based on the number of operating locations and number of people served in each location; for example, planned scale-up to 20 locations will present different issues from scale-up to only 5 locations

The need for cost-effectiveness and benefit-cost studies becomes more prominent at this stage, since the evidence will be especially helpful to convince funders of the importance of providing ongoing support for expanded operations. Since it is often impossible for one organization to serve the desired number of beneficiaries of an intervention, this is also the stage during which it may be especially important to explore approaches that go beyond an individual organization's own programmatic footprint. These broader scaling strategies might include, for example, public policy efforts, training models, and use of electronic platforms to deliver certain aspects of the program. Although changes in public policy and systems, as well as the prospect for public sector funding, should have been considered from the outset,[33] interventions at scale may be particularly able to influence — and will be influenced by — these matters. Since strategies to scale impact during this stage occur under conditions that may well differ from those in place for earlier evaluations, it will be important to re-test the impact to ensure that changes in delivery or adjustments to the program elements produce equally favorable results.

33 Fixsen, Blase, Horner, and Sugai (2009).

SECTION III
NEXT STEPS FOR THE FIELD

This Guide has reviewed evidence-building in the context of scale-up, primarily with respect to evaluating the effectiveness of programs. As private funders continue to consider scale-up investments, it will be necessary to address a number of important issues, including how best to apply the lessons and principles in this Guide to other types of interventions. In each case, the eight questions in the first section of this Guide should be asked in varying degrees, although the field of social investment needs to grapple with how to apply these questions to other scaling strategies, such as policy initiatives, practice dissemination, systems change efforts, social movements, and infrastructure projects.

For some of these strategies, such as practice dissemination, it may be feasible to conduct rigorous evaluation studies with counterfactuals and randomly assigned control groups. For others, such as policy and systems change, an impact study may be more challenging and less definitive in its findings. However, even in these situations, funders should consider how evidence of impact can be ascertained before policies or systems change efforts are scaled. For example, systems change should both rest on and support strong underlying program components that have rigorous evidence of effectiveness.

The Social Impact Exchange has begun to codify some of the elements needed to assess these scale-up strategies, including in drafts of due-diligence frameworks on "Deploying a New Product or Platform" and on "Scaling Policy Initiatives."[34] Other organizations that can provide insights on these issues include the Innovation Network[35] and the Alliance for Justice.[36] In addition, there are examples of evaluations that examine the impact of "place-based" interventions on neighborhood or other community outcomes.[37]

While there is much that needs to be learned about how to apply questions in this Guide to different types of scaling strategies, there is no doubt that assessing the best evidence for scale-up is a critical piece of the decision-making process. This Guide should serve as fertile ground for future studies and action guides. As others have pointed out,[38] key steps to strengthen the evidence-building process will need to include:

- Candid acknowledgment when there are limitations in the reliability of evidence to guide decision-making about scaling.
- Assurance that sufficient resources are provided to collect quality data, to interpret the data, and to communicate the evidence in usable form to key decision-makers. Funders cannot expect evidence of effectiveness to be available if they do not invest in the evidence-building process and openly share their findings, including those that are not positive.
- Patience in the cumulative process of evidence-building, as well as careful thought about when less evidence might be needed for certain decisions and about how best to balance the desire for rigorous evidence with the realities of limited resources to develop the evidence.

---

34 See Social Impact Exchange (2013).
35 See the Innovation Network (2013) Web site (www.innonet.org).
36 See the Alliance for Justice (2013) Web site (www.allianceforjustice.org).
37 Galster, Temkin, Walker, and Sawyer (2004).
38 See, for example, Bugg-Levine and Emerson (2011) and Kramer, Parkhurst, and Vaidyanathan (2009).

- Agreement on common definitions of outcomes and how they will be measured.
- Improved methods to understand and specify the key elements that contribute to a program's positive results.
- High-quality data on program implementation and participant outcomes to improve performance continuously during scale-up, perhaps drawing from approaches used in the for-profit sector.
- Attention to the public policy and systems changes needed to support implementation of effective interventions at scale.

Challenging as this will be, it can be achieved if there is a commitment by private and public funders alike to a continuous and collaborative process of building evidence to inform scale-up decisions.

APPENDIX
STAGES OF SCALE-UP DECISIONS

	EARLY→ DEVELOPING (PRE-SCALE-UP)	DEVELOPING→ PROMISING (PRE-SCALE-UP)	PROMISING→ EFFECTIVE (LIMITED SCALE-UP/ REPLICATION MAY BE APPROPRIATE AS PART OF LEARNING AGENDA)	EFFECTIVE→ SCALING (SCALE-UP RANGING FROM MODEST TO EXTENSIVE IN # OF LOCATIONS AND SIZE OF OPERATIONS IN EACH LOCATION)
PRIORITY EVALUATION CONSIDERATIONS	• Understand the needs and opportunities of the population being served and how they are selected *(Questions 1 and 2)* • Conduct a careful review of local conditions, operating lessons, and results of evaluations of similar programs for similar populations *(Question 1)* • Explore alternative approaches, develop a theory of change and program model, and consider early implementation lessons *(Questions 1 and 3)* • Track participation characteristics, early performance measures (especially participation levels in key program components), and outcomes *(Questions 2, 3, and 8)*	• A funnel analysis reflecting steps in the recruitment and selection process, the extent to which individuals progress through the steps, and the reasons for any drop-offs, especially whether program staff excluded certain people considered more difficult to serve *(Question 2)* • Further development of the theory of change, based on early operating experience *(Question 3)* • Documentation that threshold levels of participant engagement have been reached • Establishment of specific outcome goals that are reasonable in light of the participant characteristics and selection process; documentation of progress in achieving the outcome • Initial cost estimates, but not yet cost-effectiveness or benefit-cost figures	• Testing operations more broadly in the regular service delivery system • The extent to which the program *causes* improved outcomes, using random assignment or strong quasi-experimental design *(Question 4)* • Complementary implementation and cost analyses, perhaps including cost-effectiveness or benefit-cost studies *(Questions 3, 5, and 6)* • Assessment of fidelity to the key elements of the model, although model refinement continues *(Questions 3, 4, 5, and 7)*	• The effect of scale-up on participant characteristics *(Question 2)* • Impact and implementation analyses, focusing on maintaining or improving results in context of scale-up, potentially across numerous locations *(Questions 3, 4, 5, and 7)* • Cost-effectiveness and/or benefit-cost analysis *(Question 6)* • Refined analysis of appropriate balance between fidelity and local variation • Plans for continued evaluations

	EARLY→ DEVELOPING (PRE-SCALE-UP)	DEVELOPING→ PROMISING (PRE-SCALE-UP)	PROMISING→ EFFECTIVE (LIMITED SCALE-UP/ REPLICATION MAY BE APPROPRIATE AS PART OF LEARNING AGENDA)	EFFECTIVE→ SCALING (SCALE-UP RANGING FROM MODEST TO EXTENSIVE IN # OF LOCATIONS AND SIZE OF OPERATIONS IN EACH LOCATION)
HOW FUNDERS CAN HELP	• Connect grantees with key constituencies (including the people whom the intervention will serve) and experts in the field to understand the local context and lessons from previous evaluations • Ensure that the information collected is synthesized and used to help sharpen the theory of change and to develop the program model • Encourage operating flexibility and experimentation • Support development of a performance tracking system (with unique participant identifiers) that is tied to the theory of change • Begin to understand potential contributions of specific program components	• Support improvements in the participant tracking system, including data needed for a funnel analysis and continuous program improvement (for example, ways to increase participation in key program components) • Possibly fund independent evaluator for formative, implementation, or outcome analysis; net impact studies typically not conducted yet	• Support impact, implementation, and cost evaluations by independent, third-party evaluator(s) • Support analysis of which components should be implemented with fidelity and which are better left to local adaptation • Support dissemination of evaluation findings • Convene potential funding partners • Provide capacity-building grants to prepare for scale-up • Provide capacity-building grants to strengthen data collection and analysis for continuous quality improvement	• Support impact, implementation, and cost evaluations by independent, third-party evaluators; include cost-effectiveness and/or benefit-cost studies • Support development of a "fidelity guide" clarifying elements to control and elements for which variation is appropriate • Provide capacity-building grants to sustain scale-up • Provide continued support for dissemination of evaluation findings • Continue to help convene funding partners

	EARLY→ DEVELOPING (PRE-SCALE-UP)	DEVELOPING→ PROMISING (PRE-SCALE-UP)	PROMISING→ EFFECTIVE (LIMITED SCALE-UP/ REPLICATION MAY BE APPROPRIATE AS PART OF LEARNING AGENDA)	EFFECTIVE→ SCALING (SCALE-UP RANGING FROM MODEST TO EXTENSIVE IN # OF LOCATIONS AND SIZE OF OPERATIONS IN EACH LOCATION)
WHAT'S NEEDED TO MOVE TO THE NEXT STAGE	• A plausible theory of change that is grounded in the available research and in the early operating experience; it should show potential pathways to success, beginning with how participants are recruited and enrolled and extending through the mix and sequence of services that are expected to produce the desired outcomes • A reasonably well-developed participant tracking system • Initial anecdotal and other evidence suggesting that the program is feasible to operate and has genuine potential to produce the desired results	• A better understanding, based on early operating experience, of the organizational, systems, and policy context for the program • Understanding of participant characteristics (for program design and interpreting outcomes) • A reasonably well-developed theory of change • A functioning internal system to capture participant characteristics and to track participant activity and outcomes • Positive internally generated operating and outcome data demonstrating that participation levels and outcomes are consistent with a theory of change • Acceptable operating costs	• Positive net impact results *(Question 4)* • Sound theory of change and clear articulation of key program elements • Strong internal system to track participant characteristics and activity levels, staff activity, and program outcomes in context of program expansion and/or replication • Reasonable fidelity across program sites and quality-control measures to maintain fidelity • Strong support for scale-up within the operating organization(s), system of services, and policy/funding environment *(Question 1)*	• Calibration of the rigor of evaluation and corresponding investment based on the number of operating locations and number of people served in each location; for example, planned scale-up to 20 locations will present different issues from scale-up to only 5 locations

REFERENCES

Baron, Jon. 2012a. *Rigorous Program Evaluations on a Budget: How Low-Cost Randomized Controlled Trials Are Possible in Many Areas of Social Policy*. Washington, DC: Coalition for Evidence-Based Policy.

Baron, Jon. 2012b. *Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?: A Brief Overview and Sample Review Form*. Washington, DC: Coalition for Evidence-Based Policy.

Bloom, Howard S. 2012. "Modern Regression Discontinuity Analysis." *Journal of Research on Educational Effectiveness* 5, 1: 43-82.

Bloom, Howard S. 2005. *Learning More from Social Experiments: Evolving Analytical Approaches.* New York: Russell Sage Foundation.

Bloom, Howard, Carolyn Hill, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness* 1: 289-328.

Bugg-Levine, Antony, and Jed Emerson. 2011. *Impact Investing: Transforming How We Make Money While Making a Difference.* San Francisco: Jossey-Bass.

Cellini, Stephanie Riegg, and James Edwin Kee. 2010. "Cost-Effectiveness and Cost-Benefit Analysis." In Joseph Wholey, Harry Hatry, and Kathryn Newcomer (eds.), *Handbook of Practical Program Evaluation.* Third Edition. San Francisco: Jossey-Bass.

Coffman, Julia. 2010. "Broadening the Perspective on Scale." *The Evaluation Exchange: A Periodical on Emerging Strategies in Evaluation* 15, 1.

Duggan, Anne, Debra Caldera, Kira Rodriguez, Lori Burrell, Charles Rohde, and Sarah Shea Crowne. 2007. "Impact of a Statewide Home Visiting Program to Prevent Child Abuse." *Child Abuse & Neglect* 31, 8: 801-827.

Durlak, Joseph A., and Emily P. DuPre. 2008. "Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation." *American Journal of Community Psychology* 41: 327-350.

Fixsen, Dean L., Karen A. Blase, Rob Horner, and George Sugai. 2009. *Scaling Up Evidence-Based Practices in Education*. Chapel Hill, NC: FPG Child Development Institute, The University of North Carolina at Chapel Hill.

Galster, George, Kenneth Temkin, Chris Walker, and Noah Sawyer. 2004. "Measuring the Impacts of Community Development Initiatives: A New Application of the Adjusted Interrupted Time Series Method." *Evaluation Review* 28, 6.

Gueron, Judith. 2005. *Throwing Good Money After Bad: A Common Error Misleads Foundations and Policymakers*. Stanford, CA: Leland Stanford Jr. University.

High/Scope Educational Research Foundation. 2012. *The Youth Program Quality Assessment: A Research-Validated Instrument and Comprehensive System for Accountability, Evaluation, Program Improvement.* Ypsilanti, MI: High/Scope Educational Research Foundation.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142: 615-635.

Kemple, James J. 2008. *Career Academies: Long-Term Impacts on Work, Education, and Transitions to Adulthood*. New York: MDRC.

Kramer, Mark, Marcie Parkhurst, and Lalitha Vaidyanathan. 2009. *Breakthroughs in Shared Measurement and Social Impact*. Boston: FSG Social Impact Advisors.

Major, Dara. 2011. *How Do We Approach Impact and Evaluation in the Context of Scale? Reframing the Conversation: A Geo Briefing Paper Series on Growing Social Impact*. Washington, DC: Grantmakers for Effective Organizations.

McDonald, Sarah-Kathryn. 2010. "Developmental Stages for Evaluating Scale." *The Evaluation Exchange: A Periodical on Emerging Strategies in Evaluation* 14, 1.

Pianta, Robert C., Bridget B. Hamre, Jason T. Downer, O. Barbarin, D. Bryant, and C. Howes. 2008. "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic Language and Social Skills." *Child Development* 79: 732-749.

Preskill, Hallie, and Tanya Beer. 2012. *Evaluating Social Innovation*. Washington, DC: Center for Evaluation Innovation.

Reisman, Jane, Anne Gienapp, Kasey Langley, and Sarah Stachowiak. 2004. *Theory of Change: A Practical Tool For Action, Results and Learning.* Seattle: Organizational Research Services.

Rutnik, Tracey A., and Marty Campbell. 2002. *When and How to Use External Evaluators*. Baltimore: Association of Baltimore Area Grantmakers.

Seftor, Neil S., Arif Mamun, and Allen Schirm. 2009. *The Impacts of Regular Upward Bound on Postsecondary Outcomes 7-9 Years After Scheduled High School Graduation*. Princeton: Mathematica Policy Research, Inc.

Siegel, Joanna E., Milton C. Weinstein, Louise B. Russell, and Marthe R. Gold. 1996. "Recommendations for Reporting Cost-Effectiveness Analyses." *The Journal of the American Medical Association* 276, 16: 1339-1341.

Social Impact Exchange. 2013. *Knowledge Center Tools and Templates*. New York: Social Impact Exchange. Web site: www.socialimpactexchange.org/exchange/knowledge-center/tools-and-templates.

Twersky, Fay, Jodi Nelson, and Amy Ratcliffe. 2010. *A Guide to Actionable Measurement*. Seattle: Bill & Melinda Gates Foundation.

W. K. Kellogg Foundation. 2004. *Logic Model Development Guide*. Battle Creek, MI: W. K. Kellogg Foundation.

Weiss, Heather B., Julia Coffman, Erin Harris, Priscilla M. D. Little, Heidi Rosenberg, Helen Janc Malone, and Katie Chun. 2010. "Scaling Impact." *The Evaluation Exchange: A Periodical on Emerging Strategies in Evaluation* XV, 1: 1-23.

Weiss, Michael, Howard A. Bloom, and Thomas Brock. 2013. *A Conceptual Framework for Studying the Sources of Variation in Program Effects*. New York: MDRC.

Social Impact Exchange
TAKING SUCCESSFUL INNOVATION TO SCALE

Social Impact Exchange
at Growth Philanthropy Network
122 East 42nd Street, 17th floor
New York, NY 10168
Tel: 212 551 1148

www.socialimpactexchange.org

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

MDRC
16 East 34th Street
New York, NY 10016
Tel: 212 532 3200

Regional Office
475 14th Street
Oakland, CA 94612
Tel: 510 663 6372

www.mdrc.org