THE BIG WHY

A Learning Agenda for the Scale-Up Movement

BY ROBERT C. GRANGER

Policymakers and practitioners who believe that research evidence should inform policy and practice face several challenges. These include debates about the standards of evidence for allocating resources to programs, weak information on how to produce change at scale, and concerns that a few, well-evaluated programs will drive out others that deserve support. Such challenges threaten to undermine 30 years of progress in learning which social programs improve child, youth, and family outcomes. The purpose of this article is to describe a strategy that can inform these and other issues facing evidence-based policymaking.

Take, for example, programs and policies aimed at improving the well-being of young people. The standard evidence-based position assumes widespread improvement for children and youth will occur through "scaling-up" brand-name programs, models, and organizations that have produced effects in prior evaluations. Do more of what works and less of what does not; the idea seems prudent and has political appeal. There is currently great interest in this approach in the public sector, fueled in part by the availability of federal stimulus funds geared toward scaling up evidence-based programs. Examples include the White House's Social Innovation Fund (SIF), the Department of Education's Investing in Innovation Fund (i3), and funding from the Department of Health and Human Services to replicate evidence-based home visitation and teenage pregnancy prevention programs. These initiatives are bold in scope and in their commitment to doing what works.

Prior history shows that programs that are effective at small scale have trouble maintaining that effectiveness when replicated more broadly. Recognizing this, the new initiatives include funding to support building the capacity of existing organizations to implement the evidence-based programs and, for larger projects, strong

evaluation designs to test the effectiveness of the program at scale. This is fortunate because it creates a foundation for providing guidance on questions for which we currently have no conclusive answers: (1) What policies and other conditions improve the likelihood that programs will have positive effects? (2) What organizations or other program-level policies and practices lead to positive effects?

Much research and development work is focused on clarifying the effects created by schools, youth organizations, and programmatic interventions. My argument is that too little of this work examines the conditions, policies, and practices that produce such effects. In today's vernacular, we need more research attention paid to *why* and *under what conditions* things work as the missing ingredients in the "what works" agenda.

The good news is that the launch of the various federal initiatives creates an exceptional opportunity to improve our answers to these *why* and *when* questions. Understanding the answers to these questions would improve our ability to expand effective programs in a way that maintains their effectiveness. Using the new initiatives to pursue these questions has the added advantage of leveraging them to more effectively justify their cost in the current fiscal environment. We will learn about the effectiveness of this work, while also gaining enough knowledge to do even better work the next time. It is an opportunity we should not waste. Before describing how policymakers might pursue the learning agenda, I will explain why I am concerned.

Scale-Up in Practice

For the past seven years, I have been president of the William T. Grant Foundation. Part of running a mid-sized foundation strategically is operating in a way that is flexible and complements the work of larger public and private funders. Given our focus on vulnerable youth, those funders include research agencies such as the National Institutes of Health (NIH) and the Institute of Education Sciences (IES), as well as private funders such as the Edna McConnell Clark Foundation, the Spencer Foundation, and the Bill & Melinda Gates Foundation.

Historically, we, along with our colleagues, have pursued scale-up strategies as we tried to improve outcomes for vulnerable children, youth, and families. One version of scale-up assumes that researchers will develop and incubate new strategies or programs, test those programs under limited circumstances, and then work with policymakers and practitioners to implement and test them at scale. This approach is rooted in the tradition of phased clinical trials in medicine, and NIH and IES favor it. The development of David Olds's Nurse-Family Partnership is a good example, and congressional staff referenced that program heavily when the decision was made to scale-up home visitation as part of health care reform.

A closely related strategy, perhaps best exemplified currently by the Edna McConnell Clark Foundation, is to search for promising organizations, encourage strong evaluations of organizational impact, and then expand the organizations that have promising evaluation results. This approach is similar to the strategy businesses use to expand their services and market share. Not surprisingly, it is advocated by many of the management consulting firms that are currently working with philanthropic organizations. While NIH has funded many good evaluations of researcher-created programs, there are fewer strong studies of practitioner-developed programs, in part because many organizational leaders have avoided strong tests of their organizational impact. Yet, there are examples—the BELL Accelerated Learning Summer Program (BELL Summer) and the Carerra Adolescent Pregnancy Prevention Program are two.

The two scale-up approaches share a commitment to strong research and evaluation as the basis for assessing promise. This work has led to the identification of model programs and organizations that are effective at small scale, many of which are cataloged on websites created and maintained by public agencies and some nonprofit organizations. The most ambitious example of such a site, and perhaps the best, is the What Works Clearinghouse (http://ies.ed.gov/ncee/wwc/) sponsored by the federal Department of Education. Other prominent examples include the Coalition for Evidence-Based Policy's Social Programs That Work (http://evidencebasedprograms.org/wordpress/), Johns Hopkins University's Best Evidence Encyclopedia (http://www.bestevidence.org/), and the University of Colorado's Blueprints for Violence Prevention (http://www.colorado.edu/cspv/blueprints/).

Concerns about the Scale-Up Model

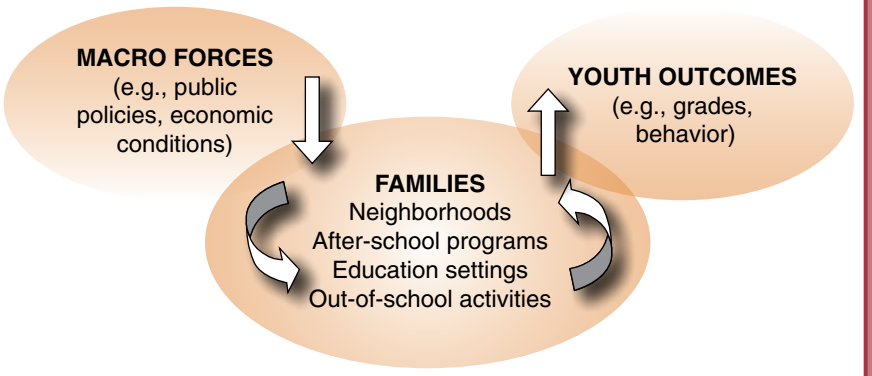
Despite the research community's ability to identify promising programs, there is almost no evidence that it is possible to take such programs to scale in a way that maintains their effectiveness. A recent report from the National Academies underscores this concern.

The 2009 report *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities* concludes that substantial progress had been made in identifying efficacious interventions during the past 15 years, but that "thus far, however, preventive interventions have not been widely implemented in schools and communities and have done little to reduce behavioral health problems in American communities" (p. 297). While calling for more research on how to "implement and disseminate" interventions, the report also quotes a paper by Dean Fixsen and colleagues that synthesized what is known about the problems of implementation and replication of model programs. Fixsen and colleagues argue that "successful implementation is synonymous with coordinated change of system, organization, program, and practice levels," and note that such coordination rarely exists.

Most current scale-up initiatives, including those the Obama administration is launching, are consistent with the Fixsen analysis: Better support, incentives, and infrastructure will lead to wider diffusion of model programs and organizations. Such improvements may lead to better results. However, the mixed

success of prior efforts sends a strong message that changes via replication of evidence-based programs may never be enough to produce widespread improvements for vulnerable youth without additional adjustments to the strategy.

FIGURE 1 The ecology of youth development.



Programs as One Influence on Youth

No one is satisfied with the current outlook for youth in the United States. Too many young people lack the skills necessary for achieving success in school, work, and life. As we try to improve outcomes by increasing the availability and number of effective programs, it might be useful to consider how such programs fit into the larger array of forces that affect young people. Figure 1 depicts how youth development is influenced by what happens in the daily environments where youth spend their time: classrooms, households, neighborhoods, community-based programs, and in informal activities with peers and others. What happens in any one of these daily settings is influenced by what happens in the others (e.g., events at home influence what goes on in school and vice versa).

Powerful secular trends and unpredictable historical events shape these daily settings, as do public policies. For example, shifts in immigration patterns alter who is in our classrooms, an oil spill affects household incomes, and the evolving labor market influences how much the skills developed in a youth employment program are rewarded in the job market. Similarly, policies shape the nature of programs in intended and unintended ways (e.g., changes in accountability policies are meant to improve what goes on in schools but may also encourage more test preparation in lieu of other teaching).

Figure 1 is a useful reminder that we ought to be modest in our expectations of any scale-up effort that does not transform daily life, and programs are unlikely to be as transformative as the policies, secular trends, and historical events that shape youth and their daily settings. This makes it all the more impressive when evidence-based programs do beat the odds and make a difference for young people. (The criteria used in the social sciences to confer "evidence-based" on a program requires that it produce improvements in youth outcomes greater than those that would have happened without the program.)

Documenting this difference is not as difficult as it may seem. The best evaluation designs for measuring program effects (known as field experiments) use a lottery to allocate access to a program when it has excess demand. The lottery creates two equivalent groups, one who can attend the program under study and another who can attend similar programs in the community. The two groups are followed, and when the outcomes for the two groups differ at a level not likely to be due to chance, the difference is logically attributed to the difference in experiences created by one group having access to the program of interest.

Learn When and Why Programs Are Effective

In the past 30 years, we have become much better at understanding how to conduct such lottery-based studies in "real-world" settings to produce accurate estimates of program effects. Such studies have made it possible to have a coherent discussion of what it means to be "evidence-based." However, no single study tells us much about the conditions under which a program is effective, the policies that help it produce results, the capacities that affect an organization's ability to implement an innovation, or the staff practices that directly improve youth outcomes.

If a program produces uniformly positive effects across multiple locations, these questions are less critical. However, that is rarely the case. Summaries of such program evaluations indicate that, although programs show outstanding results in some cases, most often they produce no net gain over the status quo, and occasionally, innovative programs are less effective than existing alternatives in the community.

Learning more about why and under what conditions programs are effective is possible once you have reliable estimates of those effects. In addition, you need good measures of the conditions, policies, and practices within and outside the program that *might* influence effects, along with a large number of lottery-based studies in which such measures can be used. With the data gathered from such measures, it is possible to look across the individual studies and find the conditions, policies, and practices that predict effects.

Researchers have productively applied this strategy in a number of prior efforts. For example, in 2003, MDRC's Howard Bloom and colleagues pooled the information from three large multi-site studies of the effects of welfare-to-work strategies on participant earnings. In these studies, different local

welfare offices all used lotteries to decide if welfare recipients should receive innovative (but untested) services or services as usual, creating 59 small-scale experimental studies (i.e., one per office). Bloom and his colleagues then examined whether or not the condition of the local labor market predicted the impact of the innovative services on earnings (it did).

Prior to the Bloom analysis, some argued that innovative services for welfare recipients would be more effective when the unemployment rate was low, implying available jobs for participants if the services improved their motivation and preparation for those jobs. Others thought that the welfare-to-work programs would have less effect in such an environment, given that it would be relatively easy for clients to get jobs without help. It was also possible that people receiving welfare when unemployment was low would be particularly hard to employ and therefore difficult to help.

Bloom and colleagues found support for the first theory—welfare-to-work programs did better in labor markets in which unemployment was low. In their analysis, they found that the average program increased participant earnings by $879 during a two-year follow-up, but that a 1 percentage point increase in the local labor market's unemployment rate reduced that impact on earnings by $94, with all other factors equal. While the study could not tell us why the local labor market mattered, such a finding is useful for situating such innovative programs and predicting their effects across communities.

Bloom and his colleagues also examined whether certain welfare-to-work practices were more effective than others—at least in the short term. Some were. The programs that emphasized quick job entry increased the average effect on participant earnings ($879, as noted above) by another $720, while those that emphasized basic education as preparation for work reduced the average earnings by $16. All estimates were larger than those expected to occur by chance.

Joseph Durlak and Roger Weissberg recently produced similar work in their review of the effects of after-school programs. They synthesized the results of 66 evaluations of after-school programs, looking at the effects on nine different measures of youth performance including social, behavioral, and academic performance. On average, they found positive effects on a number of important youth outcomes assessed in the different evaluations. However, a subset of programs created large effects, and many programs created no net effects beyond those of a comparison group of youth. In trying to explain these results, Durlak and Weissberg identified four characteristics common to the subset of effective programs—each had a sequenced approach, got youth actively involved in learning, was focused on a few goals, and had activities explicitly tied to those goals. The group of programs that had the SAFE characteristics (i.e., sequenced, active, focused, and explicit) created statistically significant impacts in all nine outcome categories assessed, while the cluster of programs that did not have all four characteristics had no positive effects.

As promising as this work is, it is not common, in part because investigators are limited to analyzing data originally collected in earlier studies. For example, Durlak and Weissberg were able to reliably code for the presence or absence of the SAFE characteristics, but it seems clear that such characteristics do not affect youth directly. Rather, they are in some way related to the daily experiences that young people have in programs. It is possible that the positive effects are caused by the staff practices created in SAFE programs, and thus improving certain staff practices is the best path to achieving better youth outcomes. At this point, we do not know, because almost none of the prior after-school studies generated data on staff practices at the point-of-service. Those that did collect such information did not gather comparable data in the "control" condition, so it is impossible to know how the experiences of the two groups of youth differed over time.

A Learning Agenda for the Scale-Up Movement

Currently, it appears that federal agencies will use their various scale-up initiatives to produce reliable information on whether or not individual programs produce positive effects for young people when they are extended to new participants, organizations, and communities. However, these agencies are positioned perfectly to learn more. For example, in the Department of Education's $650 million i3 fund, a large number of innovative programs—with promising but limited track records—will receive $30 million each to try to replicate their positive effects at scale in multiple locations. Given the priorities stated for i3, many of these efforts will focus on ways to improve teacher effectiveness or help failing schools. After a few years, it is likely that the evaluations will produce the usual results—each innovation succeeded in some instances, but not in others. It is possible to take a page from analysts such as Bloom et al. and Durlak and Weissberg and increase our knowledge about *why* that happened. I will outline one possible process for gaining that knowledge.

After funding decisions are made for each of the new initiatives, it is likely that federal and state funders will require a subset of grantees—probably those with larger grants—to conduct strong impact evaluations of their expansions. The funders should then foster a consensus on common data to be collected across the impact evaluations. Progress could be made with relatively little information.

The following questions are at the heart of current debates. For each question, I've added a suggested way to collect good information to form the answer. Because we are trying to predict the patterns of effects across studies (and across sites within a sample study), this information should be collected prior to the beginning of the scale-up efforts (i.e., at "baseline").

1. **How does the rigor and extent of the prior research evidence of effectiveness predict effectiveness at scale?** (Capture the rigor and extent of prior evidence in the review process.)

2. Are programs more effective with certain youth and families than others? (Gather common measures of participants across evaluations at baseline.)

3. Are certain scale-up strategies more likely than others to produce effects at scale? (Categorize the planned scale-up strategies along practical dimensions, such as how expensive and how prescriptive they are.)

4. Are scaled-up programs more likely to make a difference in some environments than others? (Capture relevant baseline information on environmental factors that might influence effects, such as the mobility of youth or the extent to which services analogous to the innovation are available in the community.)

5. Are certain program approaches more likely than others to produce effects at scale? (Categorize program strategies along practical dimensions, such as the degree to which they are highly structured, their cost, or their presumed intensity and duration of services.)

6. Are there organizational policies, capacities, or practices that predict effectiveness when an organization replicates an evidence-based program? (Capture baseline information on proxies for organizational capacity, such as the stability of funding, leadership, and line staff.)

Not all of these data will be easy to acquire. Therefore, I would encourage a disciplined process in which a few items related to these questions are measured well. While some of this will require document review or a brief survey (e.g., information on financial stability, the baseline information on participants), much of it will be accessible from the applicants' proposals (e.g., the program approach, the scale-up plan).

I understand that there is often a large difference in what is planned and what occurs and that organizations and innovations change over time in ways that may influence the effectiveness of the innovative program. That variation will be captured by local evaluators and can be used to explain results. However, such information, gathered after the fact, is not available to funders or program operators when they are making their plans and deciding on how to allocate finite resources. My suggestion is to gather additional information earlier to be used after the study is complete, in order to better understand the variation in implementation and impacts that is likely to occur within and across the various scaled innovations. How much evidence should funders require before supporting a program expansion? And what approaches to expansion produce the best results? We can learn the answers to these questions with a little effort and foresight.

My suggestions do require some cross-study planning and agreement, though not much. The Bloom et al. experience shows that it is possible for one firm (in this example, MDRC) to collect such information across multiple states and many local programs, and the Durlak and Weissberg review proves that it is possible to extract common information from disparate evaluations done by different teams. The new initiatives could provide consistent data across a large number of individual studies in many locations. This is exactly the scenario needed to permit the analyses I am suggesting.

Such coordination may produce additional benefits. Program developers frequently talk about the features that they believe distinguish their particular innovation and rarely acknowledge that there may be a set of strategies and practices common to all effective youth programs whether or not they have been rigorously evaluated. For example, in a recent compendium of observational measures of youth program quality, Nicole Yohalem and Alicia Wilson-Ahlstrom (of The Forum for Youth Investment) examined the content of nine measures that are widely used to assess effective staff practices in youth programs. Although the measures varied slightly (e.g., some measured program management practices while others did not), all of them measured six common features of staff's work with youth: (1) the supportiveness of relationships; (2) the program environment's safety; (3) the predictability of the program's structure and routines; and practices that produced (4) positive engagement, (5) positive social norms, and (6) the opportunity to build new skills. The recognition of these commonalities is shaping subsequent work in the after-school field, as we try to identify the practices that produce good results. It is the sort of information we need in all youth fields to move beyond an endless stream of model-specific impact evaluations.

Answering the Big Why

I have argued that the results from scaling-up evidence-based programs have not been encouraging, in part because we do not know the conditions that lead to positive effects or what distinguishes the practices of programs that produce such effects from those that do not. My suggestions will not provide definitive answers to these questions. At the end, we will still have correlates of impact results, and we will not know if these correlates are causal agents. However, the ability to examine how well factors such as program context, content, and practices predict youth-level effects would put us far ahead of our current level of understanding. It is difficult to create a change in a young person's experiences that has an impact on their long-term well-being. Thanks to rigorous evaluations of the effects of social programs, under some circumstances, we have found such effects. We need to use the scale-up initiatives to help us learn why.

Robert C. Granger is President of the William T. Grant Foundation.