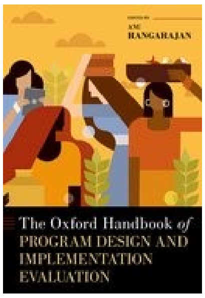


The Oxford Handbook of Program Design and Implementation Evaluation

Anu Rangarajan (ed.)

https://doi.org/10.1093/oxfordhb/9780190059668.001.0001

Published: 2023 **Online ISBN:** 9780190059699 **Print ISBN:** 9780190059668

CHAPTER

25 Effectively Scaling Up Promising Approaches: What Evidence Is Necessary and How to Ensure It Is Used to Improve Lives

Annie Duflo, Heidi McAnnally-Linz, Anu Rangarajan

https://doi.org/10.1093/oxfordhb/9780190059668.013.27 Pages 505–528

Published: 18 September 2023

Abstract

This chapter describes an approach to scale-up that focuses on delivering effective programs to increase the reach of program services. Program scale-up can occur along a number of dimensions—organizations can expand their geographic coverage, extend their time horizon, increase the number served within their existing service area, provide new services to existing clients, begin serving new groups of clients, or apply tested concepts to new problems. Similarly, programs can achieve scale by expansion or replication, and through strategic collaboration with a range of partners. This chapter lays out a framework for what types of interventions or concepts should be considered for scale-up. The chapter suggests ways to bring in measurement and research across the various stages in the process of scale-up, and discusses how to embed co-creation and ongoing learning into existing systems to enable effective scaling.

Keywords: scale-up, scaling-up frameworks, program expansion, program replication, embedding co-creation, partnerships with governments, social programs, expanding geographic coverage, ongoing learning, enabling effective scaling

Subject: Social Work

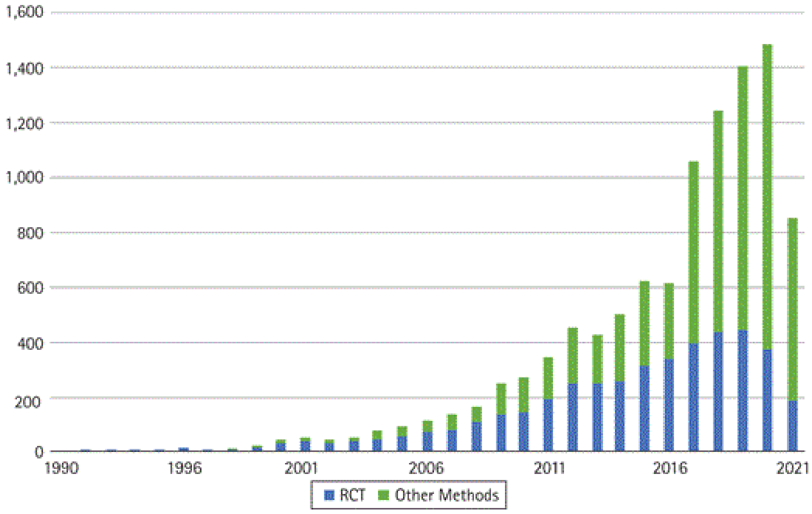
Series: Oxford Handbooks

Collection: Oxford Handbooks Online

Introduction

The global development evidence generation movement has grown in the last 20 years from a small group of pioneering researchers in the early 2000s[1] to dozens of organizations and thousands of researchers building the body of evidence and producing learnings on "what works" in a single context. For instance, as seen in figure 25.1, as of December 2022 there were more than 10,000 impact evaluations in the development sector (International Initiative for Impact Evaluation 2020). However, even though many pilots have been successful in generating impacts, relatively few have been taken to scale (McKenzie and Cull 2020). Additionally, even when interventions are expanded or taken to scale, they tend to show smaller impacts than the original study (Vivalt 2020). This chapter lays out the challenges of moving effective solutions to scale and suggests an approach to research and measurement to increase the likelihood of what works being adopted at scale and, importantly, still showing impacts at scale. ↳ p. 506

Figure 25.1



Published impact evaluations in development.

Source: https://developmentevidence.3ieimpact.org/.

1 What Does Scaling Up Mean?

The concept of scaling up development programs has been discussed since the 1970s but has become widespread with the rise of evidence-based development strategies. Scaling up is often defined as the process of "taking successful projects, programs, or policies and expanding, adapting, and sustaining them in different ways over time for greater development impact" (Hartmann and Linn 2007). According to the United Nations Development Programme (UNDP), scaling up is not just about replicating successes to cover larger groups or populations but is also about "ensuring the sustainability and adaptability of results" (UNDP 2006).

Scaling up can occur when a program is replicated or adapted to a new area or a new population, or when an existing program expands to a larger area. At the risk of simplifying, there are two common ways that an approach is scaled up:

1. **Replicating or adapting a successful pilot or effort in a new context**. This new context could be geographical or cultural (a new region or country), or it could be institutional (a nongovernmental organization [NGO] versus a government), and in some cases both. The Deworm the World initiative is

a good example of a low-cost, evidence-based approach that has been implemented in a number of countries in Asia and Africa (Evidence Action 2021).

p. 507 2. ↳ **Expanding a program to a larger geographic area.** In these instances, a program is implemented initially in one location and then scaled up to a broader area. For example, Living Goods in Uganda, following positive results, tripled the reach of their program (Björkman Nyqvist 2014). Similarly, the Ananya Family Health Initiative was first implemented in eight districts in Bihar, India, and after initial evidence of positive findings it was scaled up throughout the state (Borkum et al. 2014; Darmstadt et al. 2020).

For the purposes of this chapter, we define scaling up as the process of taking a particular program or intervention that has shown rigorous evidence of its effectiveness in a certain context (or across a number of contexts) and adapting that program or intervention to a new context and/or a much larger coverage area in the same context. Additionally, although much of the scale-up literature focuses on ensuring implementation fidelity, this is not the focus of our chapter. We assume that implementation challenges will typically exist, and measures will always need to be put in place to monitor fidelity to the core elements or concepts that are being scaled up. This chapter focuses on approaches to scaling effectively using evidence, research, and measurement, and co-creating with key stakeholders.

2 Why Is Scaling Up Important?

Ideally, implementing evidence-based programs at scale can increase the reach of effective interventions and help improve lives of the populations being served. Yet so much of the development world is built on structures and programs that are already operating at scale. Although we could see a successful pilot project that served a few thousand people and expanded to a broader geography that served a million people as a big win, the reality is that this is only a fraction of the people served daily by government agencies, schools, healthcare systems, and other entities with programs that are not necessarily evidence informed.

So how do we bridge that gap? To truly deliver on the promise of evidence to improve well-being for all, the next frontier for the evidence movement is to crack the scaling-up puzzle by learning how to work at scale with governments and implementers to improve large-scale programs with evidence.

Why hasn't the development community and ecosystem been more successful in scaling up evidence-based programs? Scaling up sounds straightforward and linear, but the challenge of scaling is much more complex. Reasons why scale-up does not occur or is not successful include the following:

- The pilot never got traction because it was implemented in a vacuum or did not consider a scale-up structure or the needs of the end users from the beginning, so the right scale-up partners were not equipped to deal with the evidence.
- p. 508 ↳ The right partners were interested (for example, the Ministry of Health or Ministry of Education), but when they tried to integrate a pilot from some other context into their existing system, it did not work at scale.
- There was full buy-in from government to test an existing program at scale, but the results did not end as hoped or expected, and the evidence was simply ignored and/or learning stopped.
- The full scope of the scale-up was under-resourced, without ongoing monitoring and learning or effective technical assistance.
- It was simply not a politically or financially viable program.

Factoring in these common reasons and addressing them in the considerations of what to scale up and how to do so may help more successfully scale up programs.

The next section provides a discussion of how to best identify what to scale up, and then describes critical research questions and measurement approaches that can facilitate successful scale-up. We end the chapter with the roles of co-creation of evidence and ongoing learning at scale.

3 What Should We Scale? Evidence Supply + Demand

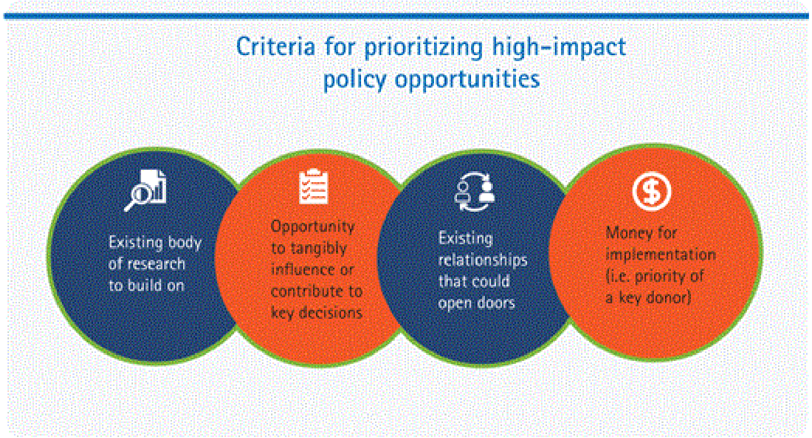
Not all programs that show impact merit consideration for scaling up, but among the pool of promising studies and programs, how do evaluators, implementers, and funders decide where to prioritize scale-up efforts? A useful framework to consider while assessing what to scale up is the generalizability framework presented by Bates and Glennerster (2017), which lays out four steps to understanding the necessary conditions for a concept to work beyond the original evaluation:

> *Step 1: What is the disaggregated theory behind the program?*
>
> *Step 2: Do the local conditions hold for that theory to apply?*
>
> *Step 3: How strong is the evidence for the required general behavioral change?*
>
> *Step 4: What is the evidence that the implementation process can be carried out well?*

Although these steps look straightforward, they can be quite challenging to answer in reality. For instance, many programs include multiple components, and the evidence may not always have identified or disentangled the relevant underlying mechanism that likely led to program success. For example, the original Pratham Balsakhi program, which implementers have tried to scale up in multiple contexts, involved a number of mechanisms, including using community tutors, identifying children who lag behind, using new pedagogies, and additional materials. It took a few iterative studies to clearly identify the key p. 509 mechanism: teaching at the level of the child. Even a seemingly ↳ simple approach, like a commitment savings product, may involve more than one mechanism—the commitment device itself, but also the fact of labeling the use of this savings account. In such cases, it is important to disentangle the mechanism to define the most cost-effective and essential programmatic elements so they can be taken to the new context.[2]

Another useful framework to consider while identifying promising scale-up opportunities is the high-impact policy opportunity (HIPO) framework developed and used by Innovations for Poverty Action (IPA), which brings together evidence supply and evidence demand in its core elements (see figure 25.2). The framework starts with a focus on carefully reviewing the existing body of research to assess when rigorous evidence is sufficient to support scale-up (first circle); it then focuses on elements that can support the demand for evidence and scaling up (the remaining three circles). Below we describe the four components of this HIPO framework as they relate to evidence supply and demand, and box 25.1 provides an illustration of a use case of these principles.

Figure 25.2


Criteria for prioritizing high-impact policy opportunities
Existing body of research to build on
Opportunity to tangibly influence or contribute to key decisions
Existing relationships that could open doors
Money for implementation (i.e. priority of a key donor)

Criteria for prioritizing high-impact policy opportunities.

3.1 Evidence Supply

The first focus of the HIPO framework is to examine the **supply of rigorous evidence** to be leveraged for building better policies and programs at scale. The team, which should include the evidence users (see section on evidence co-creation)—starts by first identifying the outcome they seek to influence (for example, learning, nutrition, income) and the context where the evidence would be applied. The next step is p. 510 to examine the whole body of evidence on a given topic or subtopic, not just single studies, and ↳ p. 511

↳ to understand which mechanisms that have already been studied in a particular topic might be most promising to scale.

Box 25.1 An Illustration of the Use of the Evidence Supply and Evidence Demand Framework on Using Growth Monitoring to Reduce Stunting in Zambia

The IPA team used the HIPO framework as we were searching for promising programs to reduce stunting, an intractable policy problem that has been prioritized by many of our key partners.

Our team identified a home-based growth monitoring program originally tested in Zambia (Muntalima et al. 2018) as promising for further testing. The program reduced stunting among stunted children by placing a life-sized growth chart on the wall inside people's homes for parents to measure the height of their children over time from 9 to 24 months old. The color-coded charts allowed parents to see immediately if a child's growth was on track or falling behind, and it also included context-specific information on nutrition. Given the relatively low cost of the growth charts, the program was also highly cost-effective. This was the first rigorous test of this concept in this particular context, but the evidence built on other results suggesting that salient reminders to encourage particular healthy behaviors could be cost-effective in improving outcomes. The team determined that the approach was worth replicating both to test whether the implementation could be done effectively via existing government health workers and also to expand to other contexts in different parts of the country where the same underlying conditions still applied (that is, a high prevalence of stunting, with parents' awareness of their child's challenges lacking).

Because the concept of growth charts might work beyond the original context, and the supply of evidence justifies its replication and measurement at scale, the effort needed was in matching evidence supply with evidence demand. As we scoped the demand for the scale-up of the growth charts program, we addressed the following three criteria:

1. **Opportunity to influence key decisions in Zambia:** IPA staff built on close to a decade of experience engaging the Ministry of Health on another program that was co-created with researchers to study how to improve the motivations of a cadre of community health workers. As the results for the original pilot growth charts program emerged, the ministry was wrapping up its own big project on stunting, for which everyone had questions regarding its effectiveness (there was no evaluation accompanying it, yet the nation's stunting numbers were not improving at the rate hoped). At the same time, IPA staff were invited to sit on the technical working group for nutrition, allowing us the platform to share evidence with all of the right stakeholders for programs to reduce stunting.

2. **Existing relationships that could open doors in Zambia:** As we shared the results with the Ministry of Health and the community health worker leadership we knew, we were introduced to the Chief Nutrition Officer, who expressed strong interest in the program but rightfully wanted more evidence on whether or not it could work in the same way at scale with government implementation. She proposed that we replicate it together. We were able to leverage this relationship along with other relationships we had with various community health worker groups, such as a collaboration with the Chief Motherhood Safety Officer, who would lead the implementation of the growth charts installation at scale, as well as with the various development partners from the technical working group, which helped build momentum and interest in the program. The Chief Nutrition Officer also joined the research team, building even stronger linkages between research and practice.

3. **Tackling a critical topic that will be prioritized and funded in the coming years:** As the interest for this program was gaining momentum through our ability to leverage the opportunity and existing relationships, the government announced an audacious goal to half stunting by 2030 and called on all development partners to support this effort, including especially UNICEF. This

prioritization ensures that this program, if proven cost-effective at this new scale with government implementation, will have substantial support to continue large-scale implementation.

The review of evidence starts with an examination of the literature, including recently published and working papers on the topic, as well as any review papers, including systematic reviews and meta-analyses, to assess if there is a clear body of evidence emerging around a particular topic. The literature and evidence reviews prioritize lessons and conclusions from randomized controlled trials, but, when relevant, they are complemented with findings from quasi-experimental studies.

Once there is a clear idea of the body of evidence on the topic, the following questions are used to identify promising mechanisms for scale:

- Was there at least one well-powered randomized controlled trial that demonstrated this particular mechanism was effective?
- How many studies have been conducted aiming to disentangle this particular mechanism? If just a few, were the underlying conditions of those contexts similar to the new context where scale-up is being considered?
- Was the other evidence on this mechanism largely positive, largely null, or perhaps mixed?
- What is the logic behind the intervention? Are the necessary conditions met in another context for this logic to apply? Should more research be considered?
- Is the promising mechanism cost-effective compared to others? Could it become more so at scale and with the right implementers?

3.2 Evidence Demand

One thing we have learned across dozens of success cases is that evidence is more likely to be used when it takes into account the needs and realities of decision-makers. However, rather than focusing on what decision makers want or what the big policy questions are, the evidence should reflect a deep understanding of the decision-making ecosystem, incentives, and opportunities. IPA's HIPO framework for scale-up prioritizes evaluators working with decision-makers to understand how evidence can be useful to them, ultimately helping "make it easy" for stakeholders to adopt the innovations (Thaler and Sunstein 2009).

p. 512 IPA's HIPO framework for scale-up focuses on three components related to **evidence demand**:

1. **The opportunity to tangibly influence a key decision** refers to using any policy or practice window, however big or small. Examples of this include working within a World Bank investment framework in the education sector, supporting an evaluation of a pilot of a government-led program that leadership is planning to scale up, informing the UNICEF strategy on early childhood education, or a large advocacy organization's strategy for reducing gender-based violence. Whatever it is, understanding the window and the ideal role of evidence in that window is a critical first step to understanding how to be useful to the donor or practitioner who might deploy that evidence.
2. **Leveraging existing relationships and understanding end users' needs** are foundational to creating change, and in particular to creating evidence-informed change. If an evaluator wants their research to be used, they must know the end users of that evidence and must work to ensure that the evidence is both useful and usable to the end users (more on how to do that in section 5). But starting a

relationship with a proposition of what a practitioner *should* be doing based on evidence almost never works. Evaluators must demonstrate their usefulness. Building an understanding of practitioner demand to become useful requires establishing relationships of mutual support between evaluators and implementers and government stakeholders. So, ideally, an evaluator can leverage their own existing relationships with practitioners or policymakers, or work closely with evidence brokers with established long-term relationships with particular institutions.

3. **Tackling a critical topic that will be prioritized and funded in the coming years** requires evidence partnerships to not only consider whom they need to influence and the current policy window, but also to take a long-term view of policy priorities. Although there may be instances of an opportunity to change a policy, as well as a strong relationship with multiple implementing organizations or ministries, a lack of funding will block scale-up. For example, a key development partner may be leaving the country and future funding around an issue may get deprioritized. Considering evidence demand requires an ecosystem approach to scale, understanding not only the politics and the people but also the flow of resources.

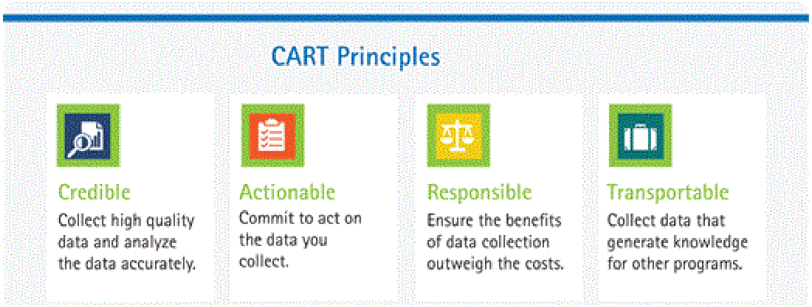
4 What Research Should We Do on the Path to Scale? Ongoing Learning to Support Scale-Up

Once a promising approach has been identified as a candidate for scale-up and all relevant stakeholders are p. 513 aligned, we need to identify the best way to scale and/or adapt ↳ the program or policy to the new context. To ensure learning along the path to scale, and ultimately the success of this process, it is critical to develop a learning agenda at the outset.

Scaling up something that worked in some other context or as a pilot program, without adequate measurement in the new context, leads to the risk of not achieving impact at scale (or knowing if the intended impacts are being achieved). Additionally, it is useful to make sure the data are useful and actionable. We recommend using the CART framework, developed by Dean Karlan and Mary Kay Gugerty (2018), as a guiding framework to build credible, actionable, responsible and transportable evidence systems (figure 25.3). This approach can help ensure that the evidence created is fit-for-purpose, and generates the right set of learnings needed across the various stages of scale-up.

The three main stages of scale-up include (1) a design and planning stage, (2) piloting and rolling out the program, and (3) ensuring implementation at scale with fidelity. Below we outline the types of research questions that will likely arise at these stages of the scaling process, and relevant measurement approaches.

Figure 25.3



CART Principles

Credible
Collect high quality data and analyze the data accurately.

Actionable
Commit to act on the data you collect.

Responsible
Ensure the benefits of data collection outweigh the costs.

Transportable
Collect data that generate knowledge for other programs.

CART principles for monitoring and evaluation.

4.1 Design and Planning

In most real-world situations, scale-up programs cannot be replicated or expanded exactly as originally implemented and will likely need some adaptation for two main reasons. First, pilots are often implemented by NGOs or other organizations, but scaled-up interventions are typically run by government or semigovernmental agents. Given that these government agents already have other responsibilities, often a larger clientele, and may or may not have the same incentives as the NGO workers, the original concept and service delivery mechanisms will likely need to be adapted. Second, adapting to the larger scale may necessitate changes to the originally implemented version of the intervention. For example, the graduation approach (Banerjee et al. 2015) for the ultra-poor was highly effective at boosting livelihoods, income, and p. 514 health among the ultra-poor ↳ across six contexts, but the unit costs were high for governments to implement the program at a large scale. Researchers are now evaluating new variations of the approach and unpacking the bundled program to understand whether the program can be made even more cost-effective.

The key types of research questions to consider during the design and planning phase while adapting a program at scale include the following:

- What is the fundamental concept or mechanisms behind the success of the original approach being scaled up?
- What are the underlying conditions in the new context? What interventions are appropriate to scale in the new context or at a large scale?
- What tweaks or adaptations should be made to the intervention elements? How much customization is needed?
- What kind of incentives for workers or beneficiaries will ensure their participation in the program? Are there any behavioral nudges that could increase the likelihood of a program being implemented?

As described below, understanding the underlying theory of change and the context of scale-up are critical elements to assessing the adaptations that need to be made in the new context.

4.1.1 Understanding and Laying out the Theory of Change

Regardless of whether the scale-up is a replication or an adapted intervention, it is critical to start with a clear understanding of the fundamental concept behind **why the original approach generated a positive impact in order to ensure those underlying conditions remain in the new context.** In other words, what is the theory of change, what are the mechanisms of impact, and under what circumstances is this approach likely to work?

The process of carefully reviewing the existing evidence supply should provide a strong foundation to understand the mechanisms driving impacts, but it is possible that gaps remain.[3] To learn more to fill the remaining gaps, it may be necessary to go back to the implementers to learn about their program and see if they can help fill gaps. Additionally, it might be useful to go to some of the original site(s) to conduct follow-up qualitative work (such as focus groups or in-depth interviews) with original program participants, as such data can help disentangle what actually happened on the ground, how people reacted to the intervention, and what they remembered of it. In the growth charts example in Zambia (box 25.1), the IPA research team went back to a number of families who received the growth charts to understand how they were using the charts. An important learning from this was that the salience and public nature of the charts was what made parents pay more attention—as a consistent reminder of the connection between healthy feeding practices and child growth and development—so the scale-up is emphasizing this learning p. 515 (Muntalima et al. 2018). ↳

Box 25.2 An Illustration of IPA's Ongoing Learning to Support Scale-Up—Ghana's Differentiated Learning Program

As IPA's presence grew in Ghana about a decade ago, we collaboratively identified all four components of a HIPO in the education sector to contextualize, replicate, and eventually scale a uniquely Ghanaian-led adaptation of the Pratham program's Teaching at the Right Level approach, now known as Differentiated Learning in Ghana.

But moving from the concept of an NGO community tutor-led model in India to a 10,000 school Ghanaian government-driven version required years of co-creation of evidence and ongoing learning. To convince ourselves (and the government) that this model could work in Ghana at scale, we rigorously evaluated four different variations of this model to answer their key questions in the Teacher Community Assistant Initiative and related evaluation (Duflo et al. 2020). The government wondered if the model could work in Ghana (it could) and whether the teachers could lead it (the union's preference and ultimately the scale-up model), or if teacher assistants were needed. Additionally, the government wondered if the model needed to be implemented during additional hours after school, or if it could be just as effective during school hours (results were similar). Although the teaching assistant model (whereby assistants were hired through the National Youth Employment Program) was more effective, it was also more costly, and ultimately the program went defunct, so the only scalable version was the teacher-led approach.

But when the results from the Teacher Community Assistant Initiative came out, though the evaluation demonstrated improved learning, it was clear that implementation fidelity mattered. And one way we expected that fidelity could improve was with adding management support. So together with the government and the whole ecosystem of actors (from the World Bank to UNICEF to local teacher training colleges, and a dozen ministry departments), we replicated again with the Strengthening Accountability to Reach All Students model (Beg et al. 2019). In short, it replicated, and learning improved because implementation of the model improved with management support.

Throughout this effort, IPA employed our co-creation tenets (see below), including co-chairing the basic education research group with the ministry, managing study steering committees with all key education stakeholders, sharing results with the right people at the right time, and advocating for resources for scale-up implementation.

In late 2019, the World Bank's new investment in the Ghana education sector was released, and it included plans to scale up the Strengthening Accountability to Reach All Students (STARS) model to 10,000 of the lowest-performing schools in the country, about half of the country or about 2 million kids. The initiative, called Ghana Accountability for Learning Outcomes Project (GALOP), was the result of years of collaborative work by dozens of organizations in Ghana's education sector. As of this writing, the teacher training for the implementation of this next phase of Ghana's differentiated learning model was underway and plans for ensuring effective scale-up via ongoing monitoring, stronger administrative data, and A/B testing of various training and delivery components of the model were being put in place to ensure effective large-scale implementation and adaptation of the model.

p. 516 Despite these efforts to understand the underlying theory of change and unpack the underlying mechanism, if gaps still remain, they can be addressed during the pilot or rollout testing phase through rapid cycle evaluations.

4.1.2 Understanding the New Context or Scale

Understanding the underlying theory of change needs to be combined with a very strong understanding of the context. What about the context is similar to or different from the previous context in which the program was originally implemented, and how might these differences be expected to affect the impacts? Detailed landscape analysis conducted via interviews with key stakeholders and focus groups with potential beneficiaries can help provide a good understanding of the context, how a potential intervention may be perceived, and what challenges or barriers may be encountered. For example, the community-led total sanitation strategy to end open defecation, originally piloted in Bangladesh, used shame and disgust as triggering mechanisms to change sanitary practices of community members. However, in a process study of a community-led total sanitation adaptation in Indonesia, we found that the adverse health effects of open defecation, particularly on child mortality, were a more powerful motivator than shame and disgust in bringing about behavior change in certain areas (Amin, Rangarajan, and Borkum 2011). In addition to studying the context, it will be important to understand the barriers to and facilitators of implementation, readiness for implementation, and other factors that could signal likelihood of success or identify elements that may need to be readied before implementation (see chapters 16 and 17 in this volume). Finally, in some instances, implementing a program on a large scale could lead to general equilibrium effects, or changes to the community at large. For instance, giving cash out at a large scale could improve local business sales or increase price inflation, or even cause psychological harm to those who don’t receive the cash; it would be useful for programs that are being considered for scale-up to assess potential broader intended and unintended consequences.

4.1.3 Laying Out Intervention Design Options and Identifying Remaining Questions

Context, combined with the theory of change, will inform several aspects of the program design, including what types of adaptations need to be made to the program or specific intervention elements, how can these be integrated into the existing programs, what needs to be customized to the current context, and so on. For example, there have been multiple evaluations of the impact of cash transfers (Innovations for Poverty Action 2015), and there is enough evidence to support the benefits they provide. However, there are questions about how to most effectively deliver cash transfers. This question became particularly important during the COVID-19 crisis, where the logistics of transporting money and identifying the right beneficiaries became an acute issue. As a result, several governments have been looking to transition to digital transfers. An example is the Ingreso Solidario in Colombia, a program created as a direct response to p. 517 COVID-19 that uses digital transfers to provide cash transfers to nonpoor but vulnerable households ↳ (Vera-Cossio et al. 2021). Similarly, in the differentiated instruction example in Ghana (box 25.2), there have now been a number of studies showing that this model using tutors or assistants leads to improved learning outcomes. However, there are practical constraints that may prevent differentiated learning from taking place in a certain context, so the question is how to best operationalize this idea to ensure that children receive differentiated learning. For example, should the assistants teach during school hours (in pull-out classes) or after school? How large should the groups be to allow for individualized attention? The answer to these questions will often depend on the context.

In addition, one size may not fit all, and the best delivery approach could vary for different population groups even within the same country or region. Going back to the differentiated instruction example, should remedial classes be held during or after school hours, and are different modalities better in different settings? A study on differentiated instruction in Ghana indicated that after-school classes could work better in smaller schools, which are generally less likely to have the space to run pull-out classes, and that during-school remedial classes were particularly effective in multigrade schools, which makes sense because these schools had more heterogeneity to start with. A carefully constructed in-depth theory of change that captures the key contextual characteristics and tries to trace the impact pathway can provide insights into which approach may work better in one context or population versus another, and also flag the assumptions that need to be tested. Good baseline data on the contextual characteristics can help those designing an intervention to allow for customization as the intervention scales. Of course, there are trade-offs between having a customized approach for population segments that may align more with a robust theory of change, and a more standardized approach, which may be easier to implement. This is itself a question that may be useful to test, depending on how much customization might matter to achieving final outcomes.

4.2 Piloting and Testing the Rollout

Although the program components, and likely the way of scaling, will be determined based on the theory of change, local context, and identifying potential operational approaches for customization and delivery, the next step is to pilot these approaches prior to rollout to validate the proposed approach and assess which is most effective. The key research questions to examine during this phase include the following:

- Which operational tweaks are appropriate to the context? Are any additional tweaks or testing needed?
- Are there any implementation challenges as these are taken to scale? What’s the best way to train a large number of implementing entities without losing the messages and ensuring consistency?
- p. 518 ↳ Is it best to provide scaling organizations a high-level blueprint that outlines the key components of the intervention (the underlying concepts that lead to impact) to adapt and customize, or to provide more specific instructions of how things should be done?
- How do we know if the interventions and the underlying mechanisms are working at scale and yielding the desired outcomes?
- What are the costs of implementing the new intervention, and what is its cost-effectiveness?

Piloting:

Piloting is a crucial step, particularly when a program that has been successfully implemented by an NGO is taken to scale in a real-world setting, and where local government agents will now be delivering services at scale. Rapid cycle evaluations and A/B testing may be particularly useful in assessing which approach is more promising in affecting proximal outcomes. These can be done quickly and use slightly less stringent criteria than a typical impact study—increasingly they can be done using existing administrative data. For instance, an A/B test could compare during- and after-school classes to track outcomes such as attendance and engagement. Additional types of piloting to understand how stakeholders perceive the intervention will also be important. For example, after having designed the Ghana replication of the India program together with the Ghana Education Service, it was piloted in a few different districts to understand how teachers would respond, test the process for recruiting the teacher assistants, assess logistical challenges with splitting classes and understanding the optimal timing for that, and so on. The pilot led to many refinements of the intervention design and made the training a lot more practical.

As the scale-up starts, the CART framework could be used to test the new model. In particular, it would be useful to examine what the right credible, actionable, responsible, and transportable approaches are to test the rollout, and assess the role of monitoring and evaluation as needed.

Monitoring:

Monitoring of intermediary outcomes to measure fidelity of implementation may be most aligned with the CART framework, depending on the existing evidence and the theory of change. The theory of change can be used to determine which intermediary outcomes are important to measure, particularly those that are critical to achieving the final outcome. For instance, in proven health interventions, such as vaccinations for immunization, or the consumption of iron-folic acid supplementation to reduce anemia during pregnancy, monitoring outcomes would focus on coverage of the vaccine or consumption of the supplements, as opposed to measuring the ultimate health outcomes that the intervention is hoping to impact. In other instances, such as the Ghana education example, where we are testing whether student performance is improving, it will be important to monitor how often the remedial classes are actually taking place, if the children attending the remedial classes are indeed the lowest-performing children, and child attendance.

An impact evaluation measuring outcomes:

Sometimes in addition to monitoring and rapid-cycle testing, a rigorous impact evaluation focused on key outcomes may be p. 519 ↳ needed. Some instances where a rigorous impact evaluation in a scale-up context will be useful are as follows:

1. **When there is still need to refine or confirm the theory of change.** For example, in some instances, the scale-up may take a proven concept or mechanism to a new place, and a rigorous impact evaluation could provide validity on the underlying mechanism that is being scaled up. For example, the key mechanism underlying the Balsakhi intervention is the TaRL (Teaching at the Right Level) approach. After the initial study, Pratham, working with the Abdul Latif Jameel Poverty Action Lab (J-PAL), developed various ways to implement and test this concept as it scaled in India, including summer camps and teacher-led differentiated learning (Banerjee et al. 2017).
2. A rigorous evaluation may be needed when it is not entirely clear how strong the dose-response relationship of the intervention is with outcomes. For example, in the education context, it may be useful to learn how often and at what frequency a child should be exposed to a remedial class for there to be an impact.
3. In the context of scale-up, it is possible that who (the entity that) delivers services may affect the outcome. Although the pilot is intended to test the new operational approach of different providers now providing services, and to ensure they have the right incentives, this part of the approach could be important to test at a larger scale as well. For example, the tutors who were volunteering for Pratham in the context of the Balsakhi evaluation may have had quite different motivations than the people who apply through the National Youth Employment Program in Ghana, which could affect the quality of their interactions with children and eventually affect learning outcomes.

Implementation study:

Monitoring and evaluation can be complemented with qualitative research—including in-depth interviews with implementers and stakeholders, focus groups with beneficiaries, and program administrative data—to understand how effectively the intervention is rolling out and if it is indeed reaching the intended beneficiaries in the way envisioned.

Measuring cost and cost-effectiveness:

The cost-effectiveness question is key to the delivery-at-scale question. It is important therefore to carefully collect cost data and to analyze these data using a cost-effectiveness framework (see chapter 22 by Long and Thornton and chapter 23 by Carter et al. in this volume).

4.3 Ensuring Fidelity and Preventing Theory of Change Drift

Often, when implementing an evidence-based program or practice, adherence or integrity to the original approach (also referred to as *fidelity*) is a key consideration. It is important to ensure that the mechanisms or components from the original approach that made it effective are adhered to in the new context, even if p. 520 adapted. As discussed ↳ earlier, scaling up in a new context will often involve tweaking a program's design and implementation strategy, so rather than ensuring fidelity of implementation, it will be important to prevent theory of change drift. That is, the core elements and underlying mechanisms will need to be maintained as the program scales up. For example, a teacher assistant program may look like the one involving the Pratham volunteers, but if the teacher assistants are not ultimately used to deliver differentiated instruction, this would not be implementing the same theory of change anymore (which identifies differentiated instruction as the pathway to impact).

Challenges to ensure fidelity of implementation at scale may vary more or less depending on the type of implementer. As discussed above, monitoring of intermediary variables is critical to ensuring effective delivery, and the indicators to monitor must be chosen based on the theory of change. In the education example, monitoring teacher assistant attendance alone is not sufficient. It will be important to monitor that the program is indeed grouping children by learning levels and delivering instruction at their level.

5 How Should We Scale? Co-creation and Ongoing Learning

The translation of evidence into programs and policies at scale is not a linear process, going from proof of concept to sharing results and to results being adopted and scaled. Similarly, scaling up what works is not an end in itself. Incorporating evidence into decision-making is a mindset, and so is incorporating decision-makers' needs into the evidence generation process. For this collective mindset to be nurtured, co-creating evidence is critical—where researchers and decision makers work together to identify and prioritize key research questions, iterate together on the design of experiments and work closely together throughout the generation of evidence and the process of learning from it. In this process, researchers work in support of decision-makers' needs.

Below, we list four tenets for researchers and their collaborators to follow as they build this mindset.

5.1 Co-creation Tenet 1: Understanding the Political and Fiscal Viability of Various Programs

It might sound obvious, but a brilliant theory-based research idea, or something that an NGO successfully implemented, can become a political nightmare at scale (and often, well before). Whether it is the teachers' union opposing teaching assistants, the primary funder wanting to focus exclusively on primary grades p. 521 when the evidence-based ↳ program pertains to secondary students, or the departure of a key political ally for a particular agenda, understanding the political landscape and fiscal viability of a particular program is critical before attempting to even pitch it to relevant scaling partners. For co-creation to work well, especially in controversial programs, it is important for the evaluation team to act as an unbiased evidence-forward advisor, and set up the plan for analysis in advance (that is create pre-analysis plans).

The research teams must also aim to understand the feasibility of various options within a particular political economy and work together with the wide variety of stakeholders key to the success of a scale-up to identify what is viable. In some instances, this may mean going with an option that a research team expects is less likely to have larger impacts at scale based on theory or another evaluation. For instance, the realities on the ground may necessitate adapting a viable program with potentially smaller impacts in order to generate some impacts nationwide, which would be better than a stronger program that never goes beyond piloting.

Working on improving an existing program carries the benefits of letting others manage the politics, and having the financing already sorted, as in the example of testing existing US Agency for International Development (USAID) programs alongside cash. This cash-benchmarking initiative by USAID has rigorously tested giving the cash equivalent of a program versus those standard programs aimed at improving malnutrition, youth employment, and smallholder agricultural profits. The role of the evaluator in this case was advising and co-creating the research questions not only with the implementing partners (such as Give Directly and Save the Children) but also with the key development partners (such as USAID), and ultimately sharing the results publicly.

Yet in other cases, as in the controversial Partnership Schools for Liberia program (Romero et al. 2020), the political will from the top to try the program existed, but the potential for scale was controversial. In this program, the government contracted out management of public schools to private providers, including some foreign for-profit providers. The president and the minister were particularly supportive of the program after visiting some of the schools the providers ran in Kenya as a novel way to dramatically improve learning outcomes in the country. But the combination of numerous donors and eight different private providers with a stake in the program, strong education advocacy groups coming out against the program, and various media groups covering the 93-school pilot made the viability of the program at scale much more questionable. Though there were clear and agreed-upon plans for the analysis and releasing the results, the evaluation team ultimately found itself releasing results that were criticized by all sides.

Finally, the need to be aware of political economy effects continues to be an issue even once an evidence-informed program is being implemented at scale, as the Y-RISE program on scaling at Yale University lays out in their political economy summary (Miquel and Finan, n.d.):

> Scaling a program can affect political behavior in a number of ways, with implications for the effectiveness of the program itself alongside other economic ↳ and political outcomes. First, externally funded aid programs may erode political accountability if ineffective leaders claim credit for successful programs (Deaton 2013). Second, large-scale programs may induce governments to reallocate effort or financial resources, potentially enhancing the effects of a program or undermining its goals.... Finally, as programs scale, they become more visible. Among constituents, this visibility can provoke public backlash, or even shift political support from one party to another, while bureaucrats and leaders may see opportunities for corruption.
>
> (Banerjee et al. 2016)

p. 522

5.2 Co-creation Tenet 2: Having Scale-Up in Mind from the Start

It is understandable to assume that the path to scale is linear: pilot, refine, test, replicate, scale. However, experience has shown that it is not that straightforward, and just providing strong results from a pilot does not automatically imply scale-up. One way to maximize chances of scaling up is to have scale-up in mind from the start. Some useful principles to keep in mind are as follows:

- Identify the right scaling partner from the beginning, and work to understand their incentives, funding, and priorities.

- Work with that scaling partner from the beginning in piloting, refining, and testing *their* ideas, informed by theory. Small NGOs might be better testing grounds for a particular mechanism, but they are unlikely to lead directly to scale.
- If it is not possible to test with scaling partners, involve them in an active steering committee (more in co-creation tenet 3).
- Consider testing only what is most scalable. A few enthusiastic leaders or trainers is not nationally scalable, so is there a way to make the program not personality dependent? A specific protocol may not be easy to follow or train trainers on—how can you simplify the intervention? Are there ways to automate?
- Consider commitment devices—can evaluators work with the scaling partner to create a pre-policy plan that details what actions they will plan to take based on the results? Are there ways to get that written into memorandums of understanding or funding and budgetary decisions?
- Work within existing systems. If the program requires an ongoing layer of management that is added to the existing system, or if the research team adds resources to an existing system that cannot be sustained over the long term, it is less likely that the program will sustain impacts at scale—design and test programs that can leverage and improve existing systems (see more in co-creation tenet 4).
- Take an ecosystem approach to evidence use by involving all critical stakeholders in co-creation, from the implementer to the funders to the local research community, even to the other civil society or advocacy groups.
- p. 523 ↳ Consider the political and financial viability of scaling the program—for example, if a major funder in a secondary education program will be leaving the country in two years, this might not be the time to try to scale a secondary program that requires more investments.

5.3 Co-creation Tenet 3: Maintaining Decision-Maker Interest and Ownership in Studies: Engagement and Capacity-Building during the Study

If the study partnership has abided by co-creation tenets 1 and 2, tenet 3 is much easier to implement, but it can also be easy to neglect. Evaluators are busy collecting and analyzing data, policymakers and practitioners are busy running programs, and funders may be on to building or implementing their next strategy. But building in mechanisms to ensure ongoing engagement and ownership over results is critical to successful scaling.

Such a mechanism can be as simple as a steering committee of key stakeholders that meets regularly to check in, share early results, visit programs to see them in action, and develop new questions to be tested. Other mechanisms include finding ways for evaluators to be useful to partners—by sharing other evidence from that context or elsewhere on how others have improved outcomes, by hosting evidence days together to rally more support for evidence, or by building evidence labs or supporting monitoring systems (see tenet 4 for more). For example, in the Zambia growth charts scale-up (box 25.1), a steering committee regularly met to review some of the qualitative work to disentangle the results, and the research team presented a full review of the rigorous evidence in reducing stunting to that committee (as well as various other small groups). And in the Ghana education example (box 25.2), each of these mechanisms was leveraged throughout the multiple large-scale studies, building full buy-in throughout the education ecosystem, ultimately leading to the scale-up (and the ongoing measurement and adaptation of that scale-up through the lab).

Finally, a critical rule of thumb for evaluators that should never be neglected: **never ever surprise an evaluation partner by sharing results publicly before discussing the findings with them**, especially when the news is not what they would have hoped for. Unfortunately, we have seen this happen all too often, sometimes derailing possibilities for evidence use. Decision-makers and policymakers should have an opportunity to hear about results, ask their questions, make their suggestions on the framing, and prepare for their own communications around the results before anything is shared publicly or with their stakeholders. Of course, the pre-policy plan or co-creation memorandum of understanding should detail that the evaluator is unbiased and ultimately will share the results publicly. Similarly, funders can play a key role in ensuring unbiased sharing. Having laid such groundwork and having agreements beforehand on sharing results can turn bad news and a defensive reaction into an opportunity for evidence-informed ↳ change in most circumstances. Whereas examples that have gone wrong are not for public consumption, a shining example of this kind of evidence-informed change going well, bolstered by the stance of an open, learning-focused organization, is the way Sabre Trust changed their programming and continued to adapt and evaluate new versions of their program after results showed their teacher training program in Ghana was not as effective as hoped (Innovations for Poverty Action 2018).

5.4 Co-creation Tenet 4: Embedding Evidence Creation and Use within Existing Systems

In an ecosystem where evidence is used to continually improve lives, co-creation and ongoing learning do not stop with a study (or a series of studies). They are simply ways of operating and making decisions to ensure it is not just data or evidence for its own sake but credible, actionable, responsible, and transportable data and evidence (Gugerty and Karlan 2018). But this ongoing use of data and evidence requires scaling institutions to be able to continuously monitor and evaluate their own programs and make ongoing improvements.

A vehicle some organizations have employed to successfully build this culture of evidence-informed scaling and decision-making at an institutional level is the embedded labs approach—essentially embedding technical staff within an institution to help improve existing administrative data and leverage such data for rigorous research that answers critical questions for decision-making and scaling up. For example, in Ghana’s Ministry of Education, IPA staff are embedded in the Planning, Budgeting, Monitoring, and Evaluation unit, with a remit to (a) strengthen the delivery of education services by building robust learning plans and monitoring systems and data for accountability within the Ghana education system; (b) collaborate on capacity strengthening of Ministry of Education staff in data generation and use, as well as research and evaluation processes; (c) be thought partners in developing rigorous evaluations around policy-relevant aspects of proposed education programs or reforms within the next five years; and (d) provide credible empirical evidence from research programs to inform policy decisions. This unit is currently providing technical assistance to build out the monitoring of the 10,000-school scale-up of differentiated instruction and to test remaining scale-up questions along the way.

Embedded labs are most effective when they are deeply rooted in the organizational programs but operate as a part of the regular planning and budgeting teams, so that evidence is more likely to be used. This might mean that in a Ministry of Education, the lab sits either within multiple operational departments or within the planning and budget unit, but with a remit to support across other departments. In short, the lab needs to be able to access political leadership and funding partners to influence high-level decisions and funding flows, but also needs the buy-in of the implementing departments to ensure daily operations can actually improve using feedback loops. Lab staff can also serve as key connectors with external academics and conveners for ecosystem-level research or technical working groups.

p. 525 Beyond an individual institution at the country level, scaling across contexts requires an institutionalization of evidence creation and use as well. Some successful examples of this include major NGOs, like the International Rescue Committee with their Airbel lab, or BRAC University. Others are global multi-institutional initiatives such as scaling up Pratham's Teaching at the Right Level model across Africa or the Partnership for Economic Inclusion's agenda to scale up the graduation model.

To grow these successes and measurably improve many more millions of lives will require bilateral and multilateral institutions to support these initiatives not only in the institutions they fund but also in their own institutional decision-making. For example, USAID's Development Innovation Ventures unit has funded some of the most well-known evidence-informed scale-ups; however, the vast majority of USAID's expenditures are still not evidence informed. Many others have written eloquently on how to improve that (Estes et al. 2021), but one clear way is empowering a highly capable evidence unit with a clear mandate for informing the planning and budgeting processes.

6 Conclusion

Scaling up is challenging for a number of reasons, but to maximize the chances of successful evidence-informed scale-up, we need to start with matching evidence supply and demand, and build in a lot of learning along the way. But most importantly, the development community should view scaling up effective programs as a long-term co-creation process, where evaluators, policymakers, practitioners, and funders work closely together, aligning their incentives to ultimately measurably improve lives at scale.

There are many reasons evidence-based scaling can fail (see box 25.3). Co-creation can help mitigate the chances of scale-up failure, but for co-creation to work, it is important that the process is geared toward learning, as opposed to accountability, and is thought of as evidence generation, as opposed to evaluation. It should be forward looking (What should we do next?) rather than backward looking (How was your performance?). Accountability should be based on delivery, on the willingness to change based on impact results, and experimentation and learning should be rewarded, even if impact results are negative.

For evidence-informed scaling to truly deliver on the promise of evidence, each actor in the ecosystem has a key role to play (see chapter 26 by Natarajan and Zwane in this volume). Evaluators must continue to inform programs with theory and rigor but also recognize the need to co-create with scaling partners and to support the broader ecosystem in which they operate with the data and evidence that are needed for decisions. Policymakers and practitioners must commit in advance and be open to changing based on evidence. And funders must commit to flexibly funding these long-term partnerships and bridge builders,
p. 526 encouraging learning and demanding evidence for effective scaling. ↳

Box 25.3 Why Does Evidence-Based Scaling Often Fail?

Throughout this chapter, we have emphasized a process of evidence-informed scaling that works through co-creation of evidence and partnerships for ongoing learning throughout the scaling process. But many challenges can trip up evidence-based scaling, and chief among them is approaching an evidence-based program as if it is evidence itself that is scaling, rather than a program that may scale (or not) based on many factors, including evidence.

Here are a few ways evidence-based scaling can fail:

1. Politics → public perceptions, unions, donors, elections.
2. The speed of change.
3. Programs from outside being imposed on a system, rather than integrated in.
4. Co-creation process breaks down because researcher/implementer incentives + time don't necessarily match.
5. Donor restrictions combined with output-based funding that focuses on the numbers of people reached or number of trainings delivered don't lend themselves to the iterative and ongoing nature of actually learning and improving.
6. Implementers don't have the ability to continue with ongoing learning once the study ends → they might need more support to build nonacademic focused learning.

Evidence-informed scaling, on the other hand, which draws on the evidence base but also takes political economy factors into consideration, encourages true partnership between those who create evidence and those who might use it to inform decisions, and understands the incentives of the scaling systems has a much higher chance of succeeding.

Acknowledgement

We would like to thank Ana Tamayo for her capable assistance, particularly with gathering the references and reviewing an earlier draft of the chapter.

Notes

[1] These included Abhijit Banerjee, Esther Duflo, and Michael Kremer, who jointly received the 2019 Nobel Prize in economics for their experimental approach to alleviating poverty in developing countries.

[2] This additional research can include qualitative approaches to disentangle what happened, replications of the same study in a different context to help confirm and refine the theory of change, and/or evaluations of additional or disentangled treatment arms.

[3] Unfortunately, too often research papers do not adequately describe the theory of change and the expected causal pathways to impact.

p. 527

References

Amin, Samia, Anu Rangarajan, and Evan Borkum. 2011. *Improving Sanitation at Scale: Lessons from TSSM Implementation in East Java, Indonesia*. Princeton, NJ: Mathematica.
Google Scholar Google Preview WorldCat COPAC

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." Working Paper 22931. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w22931.
WorldCat

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348 (6236). https://doi.org/10.1126/science.1260799.
Google Scholar WorldCat

Bates, Mary Ann, and Rachel Glennerster. 2017. "The Generalizability Puzzle." *Stanford Social Innovation Review*, Summer 2017. https://ssir.org/articles/entry/the_generalizability_puzzle.
Google Scholar WorldCat

Beg, Sabrin, Anne Fitzpatrick, and Adrienne M Lucas|. 2019. "The Importance of Management Support for Teacher-Led Targeted Instruction in Ghana." Washington, DC: Innovations for Poverty Action. https://www.poverty-action.org/sites/default/files/publications/STARS%20Baseline_Final.pdf.
WorldCat

Beg, Sabrin, Anne Fitzpatrick, and Adrienne Lucas. 2019. "*Strengthening Teacher Accountability to Reach All Students (STARS).*" Nashville, TN: American Economic Association. https://doi.org/10.1257/rct.3977-1.0.
Google Scholar Google Preview WorldCat COPAC

Björkman Nyqvist, Martina. 2014. "Entrepreneurial Community Health Delivery in Uganda: A Cluster-Randomized Controlled Trial." Nashville, TN: American Economic Association. https://doi.org/10.1257/rct.530-1.0.
WorldCat

Borkum, Evan, Anu Rangarajan, Dana Rotz, Swetha Sridharan, Sukhmani Sethi, Mercy Manoranjini, and Lakshmi Ramakrishnan, Lalit Dandona, Rakhi Dandona, Priyanka Kochar, Anil Kumar, and Priyanka Singh. 2014. "*Midline Findings from the Evaluation of the Ananya Program in Bihar.*" Final report to the Bill & Melinda Gates Foundation. Princeton, NJ: Mathematica Policy Research, December 2014.
Google Scholar Google Preview WorldCat COPAC

Darmstadt, Gary L., Yingjie Weng, Kevin T. Pepper, Victoria C. Ward, Kala M. Mehta, Evan Borkum, Jason Bentley, et al. 2020. "Impact of the Ananya Program on Reproductive, Maternal, Newborn and Child Health and Nutrition in Bihar, India: Early Results from a Quasi-Experimental Study." *Journal of Global Health* 10 (2): 021002.
Google Scholar WorldCat

Deaton, Angus. 2013. *The Great Escape*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/books/hardcover/9780691153544/the-great-escape.
Google Scholar Google Preview WorldCat COPAC

Duflo, Annie, Jessica Kiessel, and Adrienne Lucas. 2020. "Experimental Evidence on Alternative Policies to Increase Learning at Scale." Working Paper 27298. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w27298.
WorldCat

Estes, Jocilyn, David Evans, and Sarah Rose. 2021. "The Case for Evidence-Based Innovation and Implications for USAID (and Beyond)." *Center for Global Development* (blog), February 25, 2021. https://www.cgdev.org/blog/case-evidence-based-

innovation-and-implications-usaid-and-beyond.
WorldCat

Evidence Action. 2021. “Our Fight against Worms.” Washington, DC: Evidence Action. https://www.evidenceaction.org/.
WorldCat

p. 528 Gugerty, Mary Kay, and Dean Karlan. 2018. *The Goldilocks Challenge: Right-Fit Evidence for the Social Sector*. Oxford: Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780199366088.001.0001/oso-9780199366088.
Google Scholar Google Preview WorldCat COPAC

Hartmann, Arntraud, and Johannes F. Linn. 2007. “*Scaling Up: A Path to Effective Development.*” Washington, DC: International Food Policy Research Institute, October 6, 2007.
Google Scholar Google Preview WorldCat COPAC

Innovations for Poverty Action. 2015. “Cash Transfers: Changing the Debate on Giving Cash to the Poor.” Washington, DC: Innovations for Poverty Action, July 6, 2015. https://www.poverty-action.org/impact/cash-transfers-changing-debate-giving-cash-poor.
WorldCat

Innovations for Poverty Action. 2018. “Evidence-Informed Early Childhood Teacher Training in Ghana.” Washington, DC: Innovations for Poverty Action, June 13, 2018. https://www.poverty-action.org/impact/evidence-informed-early-childhood-teacher-training-ghana.
WorldCat

International Initiative for Impact Evaluation. Accessed December 2022. “*3ie Development Evidence Portal.*” New Delhi: 3ie. https://developmentevidence.3ieimpact.org/.
Google Scholar Google Preview WorldCat COPAC

McKenzie, David, and Bob Cull. 2020. “Implementing Successful Small Interventions at a Large Scale Is Hard.” *World Bank Blogs* (blog), March 19, 2020. https://blogs.worldbank.org/developmenttalk/implementing-successful-small-interventions-large-scale-hard.
WorldCat

Miquel, Gerard Padró I, and Frederico Finan. n.d. “Political Economy Effects of Policy Interventions.” New Haven, CT: Yale University. https://yrise.yale.edu/political-economy-effects/.
WorldCat

Muntalima, Claire, Peter Rockers, Doug Parkerson, and Günther Fink. 2018. “Growth Charts Project: Qualitative Follow-Up Report.” Washington, DC: Innovations for Poverty Action. https://www.poverty-action.org/sites/default/files/publications/Growth%20Charts%20Qualitative%20Follow-up%20Report_2018.12.4.pdf.
WorldCat

Romero, Mauricio, Justin Sandefur, and Wayne Aaron Sandholtz. 2020. “Outsourcing Education: Experimental Evidence from Liberia.” *American Economic Review* 110 (2): 364–400. https://doi.org/10.1257/aer.20181478.
Google Scholar WorldCat

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York: Penguin Books.
Google Scholar Google Preview WorldCat COPAC

UNDP (United Nations Development Programme). 2006. “Guidance Note: Scaling Up Development Programmes.” New York: UNDP. https://www.undp.org/content/undp/en/home/librarypage/poverty-reduction/participatory_localdevelopment/guidance-note--scaling-up-development-programmes.
WorldCat

Vera-Cossio, Diego, Bridget Lynn Hoffmann, Pablo Ibarrarán, Marco Stampini, and Camilo Jose Pecha Garzon. 2021. “The Impacts of Expanding Cash Transfers during an Emergency: Evidence from Colombia’s Ingreso Solidario Program.” Washington, DC: Innovations for Poverty Action, January 29, 2021. https://www.poverty-action.org/study/impacts-expanding-cash-transfers-during-emergency-evidence-colombia%E2%80%99s-ingreso-solidario.
WorldCat

Vivalt, Eva. 2020. “How Much Can We Generalize From Impact Evaluations?” *Journal of the European Economic Association* 18 (6): 3045–3089. https://doi.org/10.1093/jeea/jvaa019.
Google Scholar WorldCat