Stanford SOCIAL INNOVATION Review

Features

The Generalizability Puzzle

By Mary Ann Bates & Rachel Glennerster

Stanford Social Innovation Review
Summer 2017

Rigorous impact evaluations tell us a lot about the world, not just the particular contexts in which they are conducted.

BY MARY ANN BATES & RACHEL GLENNERSTER

The GENERALIZABILITY PUZZLE

In 2013, the president of Rwanda asked us for evaluation results from across the continent that could provide lessons for his country's policy decisions. One program tested in Kenya jumped out, and the Rwandan government wanted to know whether it would likely work in Rwanda as well. "Sugar Daddies Risk Awareness," an HIV-prevention program, was remarkably effective in reducing a key means of HIV transmission: sexual relationships between teenage girls and older men. A randomized controlled trial (RCT) found that showing eighth-grade girls and boys a 10-minute video and statistics on the higher rates of HIV among older men dramatically changed behavior: The number of teen girls who became pregnant with an older man within the following 12 months fell by more than 60 percent.[1]

This study was compelling partly because of its methodology: Random assignment determined which girls received the risk awareness program and which girls continued to receive the standard curriculum. Our government partners could thereby have confidence that the reduction in risky behavior was actually caused by the program. But if they replicated this approach in a new context, could they expect the impact to be similar?

Policy makers repeatedly face this generalizability puzzle—whether the results of a specific program generalize to other contexts—and there has been a long-standing debate among policy makers about the appropriate response. But the discussion is often framed by confusing and unhelpful questions, such as: Should policy makers rely on less rigorous evidence from a local context or more rigorous evidence from elsewhere? And must a new experiment always be done locally before a program is scaled up?

These questions present false choices. Rigorous impact evaluations are designed not to replace the need for local data but to enhance their value. This complementarity between detailed knowledge of local institutions and global knowledge of common behavioral relationships is fundamental to the philosophy and practice of our work at the Abdul Latif Jameel Poverty Action Lab (J-PAL), a center at the Massachusetts Institute of Technology (founded in 2003) with a network of affiliated professors and professional staff around the world.

FOUR MISGUIDED APPROACHES

To give a sense of our philosophy, it may help to first examine four common, but misguided, approaches about evidence-based policy making that our work seeks to resolve.

Can a study inform policy only in the location in which it was undertaken? Kaushik Basu has argued that an impact evaluation done in Kenya can never tell us anything useful about what to do in Rwanda because we do not know with certainty that the results will generalize to Rwanda.[2] To be sure, we will never be able to predict human behavior with certainty, but the aim of social science is to describe general patterns that are helpful guides, such as the prediction that, in general, demand falls when prices rise. Describing general behaviors that are found across settings and time is particularly important for informing policy. The best impact evaluations are designed to test these general propositions about human behavior.

Should we use only whatever evidence we have from our specific location? In an effort to ensure that a program or policy makes sense locally, researchers such as Lant Pritchett and Justin Sandefur argue that policy makers should mainly rely on whatever evidence is available locally, even if it is not of very good quality.[3] But while good local data are important, to suggest that decision makers should ignore all evidence from other countries, districts, or towns because of the risk that it might not generalize would be to waste a valuable resource. The challenge is to pair local information with global evidence and use each piece of evidence to help understand, interpret, and complement the other.

Should a new local randomized evaluation always precede scale up? One response to the concern for local relevance is to use the global evidence base as a source for policy ideas but always to test a policy with a randomized evaluation locally before scaling it up. Given J-PAL's focus on this method, our partners often assume that we will always recommend that another randomized evaluation be done—we do not. With limited resources and evaluation expertise, we cannot rigorously test every policy in every country in the world. We need to prioritize. For example, there have been more than 30 analyses of 10 randomized evaluations in nine low- and middle-income countries on the effects of conditional cash transfers. While there is still much that could be learned about the optimal design of these programs, it is unlikely to be the best use of limited funds to do a randomized impact evaluation for every new conditional cash transfer program when there are many other aspects of antipoverty policy that have not yet been rigorously tested.

Must an identical program or policy be replicated a specific number of times before it is scaled up? One of the most common questions we get asked is how many times a study needs to be replicated in different contexts before a decision maker can rely on evidence from other contexts. We think this is the wrong way to think about evidence. There are examples of the same program being tested at multiple sites: For example, a coordinated set of seven randomized trials of an intensive graduation program to support the ultra-poor in seven countries found positive impacts in the majority of cases. This type

of evidence should be weighted highly in our decision making. But if we only draw on results from studies that have been replicated many times, we throw away a lot of potentially relevant information.

MARY ANN BATES is deputy director of the Abdul Latif Jameel Poverty Action Lab (J-PAL), North America.

RACHEL GLENNERSTER is executive director of the Abdul Latif Jameel Poverty Action Lab (J-PAL).

FOCUS ON MECHANISMS

These four misguided approaches would have blocked a useful path forward in deciding whether to introduce the HIV information program in Rwanda. This is because they ignore the key insight from an evaluation: what it potentially tells us about mechanism—about *why* people responded the way they did.[4] Focusing on mechanisms, and then judging whether a mechanism is likely to apply in a new setting, has a number of practical advantages for policy making.

First, such a focus draws attention to more relevant evidence. When considering whether to implement a specific policy or program, we may not have much existing evidence about that exact program. But we may have a deep evidence base to draw from if we ask a more general question about behavior. For example, imagine a public health agency that would like to encourage health-care providers to promote flu vaccinations. They are considering whether to give providers information on how their patients' flu-vaccination rates compare with rates of their peers' patients. A review of the literature may produce few, if any, rigorous evaluations of this specific approach. The general question of how people change their behavior after learning about their peers' behavior, however, has a deep evidence base.

Second, underlying human behaviors are more likely to generalize than specific programs. Take, for example, a program in rural India run by the nonprofit Seva Mandir that one of us, Rachel Glennerster, helped evaluate. The program held regularly scheduled mobile immunization camps and, in a random subset, gave 1 kg of lentils to parents at each childhood immunization visit and a set of metal plates when the immunization schedule was completed. In communities around the incentive camps, full immunization jumped to 39 percent, compared with 6 percent in the control communities.[5] The trouble was not that parents were suspicious of vaccines. Even without incentives, 78 percent of children got at least one vaccine. But incentives helped to get parents to bring their children back regularly until the end of the schedule.

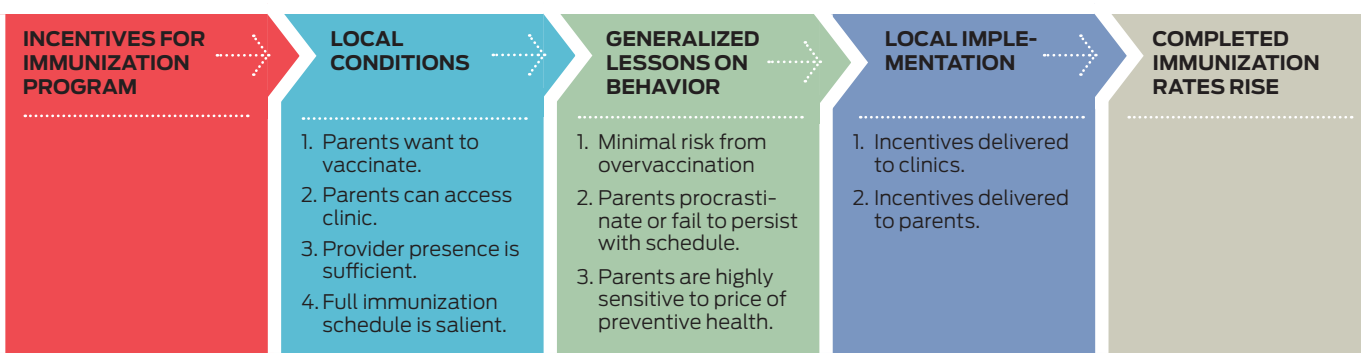
The specific program of providing lentils to encourage vaccination may not translate well to other contexts: Lentils may not be a particularly attractive incentive in other parts of the world. However, the failure of humans to maintain behaviors that help prevent future health problems generally holds: Think of all those broken diets and unused gym memberships. Similarly, the finding that the adoption of preventive health measures is sensitive to price also generalizes very well. More than half a dozen randomized evaluations of six preventive health products in five countries show that a small price cut can sharply increase demand for preventive health products.[6] Incentives can extend this finding, since they can reduce the overall cost of taking children to a clinic, which could include travel and time costs.

It is worth stressing the potentially counterintuitive point that more theory-based or "academic" impact evaluations can be particularly useful for policy purposes, because they are designed to produce general lessons. Some researchers have argued that we should have more evaluations that focus on questions that apply only to specific organizations: for example, helping Seva Mandir learn whether, locally, parents would respond better to lentils or to wheat flour.[7] But answering more theory-driven questions, such as whether take-up of preventive health is highly price sensitive, can inform the practices of many other organizations around the world.

Third, focusing on mechanisms can point us to specific local evidence that can help us predict whether a result might generalize to a new context. Common sense suggests that we are more likely to find a similar result in a new context, if the new context is similar to the one where the program was first tested. But what do we mean by "similar"? Do we mean a location that is geographically close, has the same income level, the same density of population, or the same level of literacy? There is no absolute answer. It depends on the behavior we are interested in, and it depends on theory.

What do we mean by "theory"? Theory simplifies the world to help us make (and test) predictions about behavior and about which policies are likely to be effective and where they are likely to be effective. There are many ways to make simplifying generalizations about the world. Economic theory helps us prioritize among these simplifications. For example, it suggests that what was important about giving lentils in the example above was that they are valued locally. Behavioral economic theory also suggests that people may be more sensitive to prices of preventive health than to prices of acute care when they are sick. Thus, if we want to generalize the lesson of incentives influencing the adoption of preventive health

A Generalizability Framework for Incentives for Immunization



measures, we should be more cautious if the new context focuses on acute care rather than preventive health.

The relevant theory for the immunization program also suggested that incentives would work only if parents could reliably access vaccines and were not strongly opposed to vaccines. A "similar" context therefore would be one where a large number of children got at least one vaccine (signaling the fact that access was possible and hostility to vaccines was low) but where parents failed to persist to the end of the schedule.

THE GENERALIZABILITY FRAMEWORK

At J-PAL we adopt a generalizability framework for integrating different types of evidence, including results from the increasing number of randomized evaluations of social programs, to help make evidence-based policy decisions. We suggest the use of a four-step generalizability framework that seeks to answer a crucial question at each step:

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?

To understand how this framework works, let us turn to several real-world examples of policy dilemmas. Our first case study in applying this generalizability framework concerns childhood immunizations, which are among the most cost-effective health interventions known. The World Health Organization estimates that 1.5 million more lives could be saved if immunization rates improved. Our study in India, referenced above, found that small incentives for parents, coupled with reliable services at convenient mobile clinics, increase full immunization rates six-fold, from 6 percent to 39 percent.[8] Could this approach work in Sierra Leone, which has one of the world's worst rates of mortality for children younger than 5 years old? And what about in the Indian state of Haryana or urban Karachi in Pakistan?

If we see evaluations as testing a "black box" program—if we assume that we cannot understand the mechanism at work—we would ask how many impact evaluations have tested the relationship between using incentives for immunization and immunization rates. And since only one rigorous impact evaluation assesses this relationship, we might conclude that the evidence supporting this program is quite weak. However, assessing the evidence of the different factors in the theory behind the program suggests that there is much more evidence behind this relationship than might at first be apparent.

Step 1: As we discussed above, the theory behind the original Indian study was that parents wanted to vaccinate their children—or at least had no strong opposition to vaccination. Their willingness to persist through the schedule was sensitive to small changes in price. The evidence that small costs, such as the time and transport cost of getting a child to a clinic, can deter people from persisting with preventive health behaviors is far more extensive than the black-box approach acknowledges. (See "A Generalizability Framework for Incentives for Immunization" on page 51.)

Step 2: J-PAL is working with governments in Sierra Leone; Karachi, Pakistan; and Haryana, India to determine whether the conditions required for this program hold locally. Knowledge of local institutions is important for determining basic conditions such as whether clinics open regularly and whether the vaccine supply is reliable. Publicly available data is also useful. In particular, if most children receive at least one immunization but rates fall off over the schedule, this suggests a problem similar to that observed in the original study in India. Sierra Leone, Karachi, and Haryana all fit this pattern.

Step 3: The next step concerns the evidence about behavioral conditions. Substantial evidence suggests that people worldwide underinvest in highly effective preventive health measures but spend a lot of money on acute care.[9] There is also a lot of evidence that small changes in the price of preventive health care can dramatically improve adoption rates.[10] Small incentives have also been found to have surprisingly big impacts on health behavior.[11]

Step 4: The final step focuses on the details of local implementation. Figuring out how to ensure that the incentive is delivered to the clinics and that health workers provide it to parents who get their child immunized is critical. What the incentive is, how it is delivered, and how its delivery is monitored will likely need to be adapted to the local situation.

In Karachi and Haryana, the potential to provide secure electronic payments directly to parents is dramatically reducing the logistical challenges that have plagued efforts to scale up incentives for immunization. In Sierra Leone, low penetration of mobile money in poor rural areas makes this approach less feasible. However, because of the high levels of malnutrition in Sierra Leone, agencies are keen to provide pregnant and lactating mothers with fortified food, which, local testing suggests, is highly valued. The next step in Sierra Leone, therefore, is to test whether food can be effectively distributed to parents bringing their children to be immunized. Does it reach the intended beneficiaries? Does the distribution hinder the smooth running of immunization clinics?

A Generalizability Framework for the HIV Risk Awareness Program

INFORMATION ON RELATIVE RISK OF HIV BY AGE	LOCAL CONDITIONS	GENERALIZED LESSONS ON BEHAVIOR
	1. Relationships between older men and adolescent girls are common. 2. Older men offer more financial protection against pregnancy. 3. Older men have higher rates of HIV than younger men. 4. Girls do not know that older men have higher HIV than younger men. 5. Girls trade off costs and benefits of sex with different partners.	1. Increasing perceived relative risk of HIV with one group leads to reduction in sexual activity with that group.

If the logistics of distributing food as an incentive to promote immunization proves too challenging in Sierra Leone, we should not conclude that the original study does not generalize. All we will have found is that the local implementation failed, not that the underlying behavioral response to incentives was different.

FROM KENYA TO RWANDA

We do not always proceed through every step of the generalizability framework. To illustrate this point, consider our second case study, which returns to the Rwandan government's question about preventing teenage pregnancy. How do we decide whether, in Rwanda, telling adolescent girls the relationship between men's age and HIV will help alleviate the problem? In this instance, we only used the first two steps.

Step 1: First we consider the theory behind the Kenyan HIV-information program. (See "A Generalizability Framework for the HIV Risk Awareness Program" below.) Adolescent girls trade off the benefits and costs of having sexual relationships and of having them with different partners. Girls receive various benefits from relationships with older men. In particular, older men are better able to look after them financially if they get pregnant. But relationships with older men also have risks: Older men are more likely to be infected with HIV. If girls do not know that older men are more likely than younger men to be HIV-positive, these relationships look more attractive than they really are. Knowing the relative HIV risks changes their risk-benefit calculus and reduces the number of unprotected sexual acts between teenage girls and older men.

The first steps in the theory are all assumptions about the local context, which would need to hold before we could expect that the program might work. Telling girls about the relative risk of HIV by age is not going to reduce the number of pregnancies with older men unless such relationships are common, older men have higher rates of HIV than younger men, and girls do not realize that older men have higher rates than younger men.

Step 2: The next step is to assess whether these conditions hold in Rwanda. Using publicly available data, we found that in Rwanda, too, HIV infection rates are higher among older men than younger men, and many of the teenage girls who are sexually active are so with men at least five years older than them.

But there were also important differences. In Rwanda, men ages 25-29 have an HIV rate of 1.7 percent compared with 28 percent in the district in Kenya where the original evaluation was carried out. We also found no publicly available data on perceptions of HIV risk in Rwanda. In Kenya, the fact that girls did not realize that HIV risk rose with age until they went through the program was likely to be a key driver of impact. It was therefore important to understand whether there was a similar gap between perceptions of HIV risk by age and action HIV risk by age in Rwanda.

A team from J-PAL Africa at the University of Cape Town, led by Emily Cupito, worked with the Rwanda Biomedical Center to collect local descriptive data on what teenage girls and boys knew about HIV risk. These data showed that in Rwanda most teenage girls already knew the relative risk: They correctly identified that older men were more likely to be infected with HIV than younger men. Overall, the girls in Rwanda had a pretty good understanding of the relative risk of men of different ages, although they massively overestimated the percentage of both younger and older men who have HIV. For example, 42 percent of students estimated that more than 20 percent of men in their 20s would have HIV. Only 1.7 percent of surveyed students correctly identified the HIV prevalence rate for men in their 20s as being less than 2 percent.

Note that the data that ultimately helped to diagnose whether the treatment might be effective in Rwanda did not come from an impact evaluation or an RCT. They were simple descriptive or observational data that were collected quickly (over two weeks) to assess whether the conditions were right for a program to be effective.

Funneling this local information back into our generalizability framework raised a serious concern. If an information campaign causes teenage girls dramatically to lower their perception of HIV risk associated with unprotected sex in general, but does not change their perception of relative risk, it is possible that the program could lead teenage girls to *increase* the amount of unprotected sex they have with both younger and older men.

Consequently, J-PAL did not recommend trying a "Sugar Daddies Risk Awareness" campaign in Rwanda and instead suggested exploring other mechanisms for reducing teenage pregnancy. It is important to stress, however, that we do not have a lot of evidence on exactly how and why the program worked so dramatically in Kenya. We also cannot rule out that the Kenyan program might work in Rwanda. But clearly some local conditions that theory suggests could be important for this approach do not hold in Rwanda. In this case, we concluded with the second step and recommended alternative approaches.

FROM INDIA TO CHICAGO

Depending on the mechanisms at work, lessons from one context can and do successfully transfer to other contexts. Let us turn to a final example that illustrates this point. Recently, our Education Lab colleagues in Chicago worked with the Chicago Public Schools to help high school boys who had fallen years behind the curriculum make progress. The individualized two-on-one tutoring program they tested with a randomized evaluation in collaboration with Match Education gained national attention for its large improvements in math scores.[12]

What informed the choice to try individualized learning in Chicago? The research team drew not only from quasi-experimental evaluations of high-dosage tutoring in Texas,[13] but also from randomized evaluations done in Kenya and India—contexts no one would categorize as similar to Chicago. But a look at the underlying mechanisms that helped struggling students catch up academically finds very consistent evidence across extremely dissimilar contexts.

In Kenya, an early randomized evaluation found that providing classrooms with new textbooks did not help children learn—except

A GENERALIZABILITY FRAMEWORK FOR THE HIV RISK AWARENESS PROGRAM

LOCAL IMPLEMENTATION	→	RISKY SEX WITH OLDER MEN REDUCES; RISK OF HIV REDUCES
1. Relative risk information can be conveyed effectively to girls.		

for those children who were already at the top of the class.[14] This suggested that part of the problem was that the curricula and textbooks were tailored to some but not all the wide range of learning levels in the class. A follow-up evaluation tested the idea of helping teachers provide instruction more tailored to the needs of students by grouping them by initial learning levels; it found that learning improved for students in all groups.[15]

Meanwhile, in India, Pratham, an NGO dedicated to improving education, was addressing the same challenge by enlisting local volunteers to tutor young children in basic literacy and numeracy. Though the context and approach were different, the program engaged a similar underlying mechanism: Children who had fallen years behind the official curriculum were able to catch up relatively quickly with focused teaching at the right level. Over the past 10 years, our colleagues have worked with Pratham to test different iterations of their tutoring programs in different settings: rural and urban, instruction by volunteers or government teachers, during the school day or during summer break. The results have been consistently positive.[16]

When our colleagues reviewed all the relevant evidence as they designed the Chicago study, they found parallels in the local conditions most relevant to the generalizability framework. In Chicago—as in India and Kenya—some of the students had fallen years behind the curriculum, but teachers faced incentives to teach grade-level material rather than catch-up material targeted to students' actual learning levels. Features of program implementation also had parallels: Tutors could be trained to teach to the level of the student and implement this without having to worry about managing a whole classroom with a wide range of needs. An otherwise prohibitively expensive step (teaching in very small groups) was made feasible by Match Education's approach of bringing in well-educated individuals who were willing to work for a year for a modest stipend as a public service. As in India, by removing the need for the specialized training of complex classroom management and incentives to focus only on the grade-level curriculum, they were able to run a small-group program for a more scalable cost.

This example underscores the importance of drawing connections between seemingly dissimilar studies in a way that a good literature review does. These academic reviews that discuss the common mechanisms behind effective programs are useful for policy makers precisely because they home in on the underlying behaviors that generalize across superficially different contexts. This is very different from the growing fashion in some policy circles of promoting meta-analyses, which are traditionally used in medicine and simply average the effects found across different studies. Although such meta-analysis can give an overview of a particular category of studies, it would not have helped our colleagues in Chicago: The textbook evaluation would have been averaged with other studies testing the effect of other inputs (such as chairs and desks), while the tutoring studies would have been put into another group of studies. A meta-analysis cannot draw the theoretical connections between two studies that are motivated by the same theory but test different interventions.

UNDERSTANDING CONTEXT

Too often, those who care about local context and those who do impact evaluations are seen as opposed, but this perception is false. Those of us who conduct impact evaluations and help governments integrate the lessons into policy care passionately about understanding the local context. The key to the generalizability puzzle is recognizing that we have to break any practical policy question into parts: Some parts of the problem will be answered with local institutional knowledge and descriptive data, and some will be answered with evidence from impact evaluations in other contexts.

The generalizability framework set out in this paper provides a practical approach for combining evidence of different kinds to assess whether a given policy will likely work in a new context. If researchers and policy makers continue to view results of impact evaluations as a black box and fail to focus on mechanisms, the movement toward evidence-based policy making will fall far short of its potential for improving people's lives. ■

NOTES

1 Pascaline Dupas, "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya," *American Economic Journal: Applied Economics*, vol. 3, no. 1, 2011, pp. 1-34.

2 Kaushik Basu argues that, because tomorrow is a new context, we cannot assume that a program that worked in Kenya yesterday will be effective in Kenya tomorrow. See his article "The method of randomization and the role of reasoned intuition," Policy Research working paper, no. WPS 6722, World Bank Group.

3 See Lant Pritchett and Justin Sandefur, "Context Matters for Size: Why External Validity Claims and Development Practice Do Not Mix," *Journal of Globalization and Development*, vol. 4, no. 2, 2013, pp. 161-197.

4 For a discussion of how experiments can be designed to test mechanisms, see Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan, "Mechanism Experiments and Policy Evaluations," *Journal of Economic Perspectives*, vol. 25, no. 3, 2011, pp. 17-38.

5 Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Dhruva Kothari, "Improving Immunisation Coverage in Rural India: Clustered Randomised Controlled Evaluation of Immunisation Campaigns with and without Incentives," *BMJ*, 340:c2220, 2010, pp. 1-9.

6 Pascaline Dupas and Edward Miguel, "Impacts and Determinants of Health Levels in Low-Income Countries," NBER Working Paper Series, no. w22235, 2016.

7 See Neil Buddy Shah, Paul Wang, Andrew Fraker, and Daniel Gastfriend, "Evaluations with impact: Decision-focused impact evaluation as a practical policymaking tool," 25, 3ie Working Paper, 2015.

8 For more details, see Dina Grossman, "Incentives for Immunization," *J-PAL Policy Briefcase*, November 2011.

9 See Michael Kremer and Rachel Glennerster, "Chapter Four: Improving Health in Developing Countries: Evidence from Randomized Evaluations," *Handbook of Health Economics*, edited by Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, vol. 2, 2011, pp. 201–315. Also see Pascaline Dupas, "Health Behavior in Developing Countries," *Annual Review of Economics*, vol. 3, 2011, pp. 425-449.

10 Jessica Cohen and Pascaline Dupas, "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, vol. 125, no. 1, 2010, pp. 1–45.

11 Rebecca L. Thornton, "The Demand for, and Impact of, Learning HIV Status," *American Economic Review*, vol. 98, no. 5, 2008, pp. 1829–1863.

12 Philip J. Cook, Kenneth Dodge, George Farkas, Roland G. Fryer Jr., Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg, "Not Too Late: Improving Academic Outcomes for Disadvantaged Youth," Institute for Policy Research Northwestern University Working Paper WP-15-01, 2015.

13 Roland G. Fryer, "Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments," *Quarterly Journal of Economics*, vol. 129, no. 3, 2014, pp. 1355–1407.

14 Paul Glewwe, Michael Kremer, and Sylvie Moulin, "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, vol. 1, no. 1, 2009, pp. 112–135.

15 Esther Duflo, Pascaline Dupas, and Michael Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, vol. 101, no. 5, 2011, pp. 1739-1774.

16 Abhijit Banerjee, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton, "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India," NBER Working Paper Series, no. w22746, 2016.