Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 73–102

From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application

Abhijit Banerjee, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton

Randomized controlled trials have been used in economics and other social sciences for decades. A short list of examples familiar to many economists would include the negative income tax experiments (Hausman and Wise 1985), the RAND Health Insurance Experiment (Newhouse 1993), the series of welfare reform experiments in the 1980s and 1990s (Manski and Garfinkel 1992), and work on education such as the Perry Pre-School Project and Project STAR

■ *Abhijit Banerjee is Ford Foundation International Professor of Economics and Director, Abdul Latif Jameel Poverty Action Lab, both at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Rukmini Banerji is CEO of Pratham Education Foundation and Director of the ASER Centre, both in New Delhi, India. James Berry is Assistant Professor of Economics, University of Delaware, Newark, Delaware. Esther Duflo is Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics and Director, Abdul Latif Jameel Poverty Action Lab, both at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Harini Kannan is a Senior Research Manager and Post-Doctoral Fellow, Shobhini Mukerji is the Executive Director of the South Asia regional center, and Marc Shotland is Associate Director of Training in the Research Group, all at various locations of the Abdul Latif Jameel Poverty Action Lab. Michael Walton is Senior Lecturer in Public Policy, Harvard Kennedy School, Cambridge, Massachusetts. Banerjee and Duflo are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts, and members of the Board of Directors, Bureau for Research and Economic Analysis of Development (BREAD). Their email addresses are banerjee@mit.edu, rukmini.banerji@pratham.org, jimberry@udel.edu, eduflo@mit.edu, harini.kannan@ifmr.ac.in, shobhini.mukerji@ifmr.ac.in, shotland@mit.edu, and michael_walton@hks.harvard.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.4.73

doi=10.1257/jep.31.4.73

(Schweinhart, Barnes, and Weikart 1993; Finn and Achilles 1990). Their use has accelerated dramatically in the past 10 to 15 years in academia, reflecting what Angrist and Pischke (2010) call "the credibility revolution." In terms of establishing causal claims, it is generally accepted within the discipline that randomized controlled trials are particularly credible from the point of view of internal validity (Athey and Imbens 2017). However, as critics have pointed out, this credibility applies to the interventions studied—at that time, on that population, implemented by the organization that was studied—but does not necessarily extend beyond. Some pilot studies these days are enormous, covering many millions of people (we will discuss one such study below). But in the more typical case, critics say, it is not at all clear that results from small "proof-of-concept" studies run by nongovernment organizations can or should be directly turned into recommendations for policies for implementation by governments on a large scale (for example, see Deaton 2010).

In this paper, we begin by exploring six main challenges in drawing conclusions from a localized randomized controlled trial about a policy implemented at scale: market equilibrium effects, spillovers, political reactions, context dependence, randomization or site-selection bias, and piloting bias (implementation challenges at scale). These challenges are widely recognized, and experimental evidence can often be brought to bear on them. We then turn to an example of an educational intervention called "Teaching at the Right Level" that successfully took the steps from a pilot operated by a nongovernment organization in a few slums to a policy implemented at scale by state governments in India (and in population terms, states in India are often larger than most countries in Europe). We will tell the story of how this occurred, and also how this program experienced and dealt with the six above-mentioned challenges.

While external validity of a randomized controlled trial cannot be taken for granted, is it far from unattainable. The journey from smaller-scale internal validity to larger-scale external validity is a process that involves trying to identify the underlying mechanisms, refining the intervention model based on the understanding of these mechanisms and other practical considerations, and often performing multiple iterations of experimentation.

From Proof of Concept to Scalable Policies: Six Challenges

In medical trials, efficacy studies are usually performed first in tightly controlled laboratory conditions. For the same reasons, it often makes sense to verify proof of concept of a new social program under ideal conditions—by finding a context and implementation partner where all the necessary steps for success are likely to be taken (for a formal justification of this argument, see Chassang, Padró i Miquel, and Snowberg 2012). However, the results of such a program tested on a small scale, while informative, are not necessarily a good predictor of what would happen if a similar policy were to be implemented on a large scale. Indeed, it is not uncommon that larger-scale studies fail to replicate results that had been established in small

randomized controlled trials elsewhere. In this section, we consider six obstacles that can arise in drawing conclusions from small-scale experiments, especially when the proof of concept is being taken to a larger scale.

Market Equilibrium Effects

When an intervention is implemented at scale, it could change the nature of the market. A small experiment is in many cases consistent with a partial equilibrium analysis: all relative market prices can be assumed to stay constant. By contrast, a large experiment—such as a nationwide policy intervention—is likely to affect wages and the prices of nontradable goods such as land. These price changes might affect both the overall net benefit of the program as well as the identity of the beneficiaries.

For example, a program (like a scholarship) that increases education levels for a small group will only have a minimal effect on overall education levels for the population. But as Heckman, Lochner, and Taber (1998) argue, a large-scale education intervention that produces broad increases in educational attainment across an entire population may thereby decrease the overall return to education. Thus, the results of a small randomized controlled trial of a scholarship program (as in Duflo, Dupas, and Kremer 2017) would potentially be an overestimate of the impact of scholarships on earnings, if such a program were to be scaled up.

In other settings, ignoring the equilibrium effect can lead to underestimation of the overall benefits of a treatment. For example, an intervention that increases the income among some people could lead them to consume more: if part of this consumption is in the form of nontradable goods, this will have a multiplier effect, since those who are supplying those nontradable goods will also benefit. While a small experiment may not capture this effect, it could turn out to be a source of substantial social benefits in a large-scale implementation.

An illustration of the possible pitfalls of ignoring multiplier effect is the analysis of the potential impact of microcredit. Randomized controlled trials consistently find low impact of microcredit on beneficiaries (for a recent review, see Banerjee, Karlan, and Zinman 2015). These experiments are typically based on randomization across villages, neighborhoods, or individuals. But Buera, Kaboski, and Shin (2012) suggest that microcredit may have important general equilibrium effects, and it is possible that those effects operate on a broader scale than just the village. In a nonexperimental study, Breza and Kinnan (2016) examine the sudden collapse of microcredit in Andhra Pradesh, India, following a political backlash. Contrary to the results of the previous randomized studies, they find large negative effects of losing access to microcredit and argue that this was probably the consequence of the cutback in consumption resulting from the credit withdrawal on the rest of the economy. In other words, this is a case where the general equilibrium effect is likely to be much bigger than the effect on the direct beneficiaries.

Andrabi, Das, Ozyurt, and Singh (2017) describe another mechanism for why the general equilibrium effect may be very different from the partial equilibrium effect. In their experiment in Pakistan, in some villages, one randomly selected

private school was given a grant to help improve quality. In other villages, all schools received it. The authors find very different effects on the treated schools in the two conditions. When only one school was treated, it improved its facilities at the margin and stole business from other private schools. When all schools were treated, they raised quality more substantially by investing in teachers and expanded capacity at the expense of public schools. The single-school experiment would have entirely missed this effect.

Recent research has taken this concern on board. One approach is to try to build a model to capture the various general equilibrium effects and calibrate it (as in Townsend 2010), making more- or less-heroic assumptions about the many parameters that need to be calibrated. Another approach, which has become popular, now that researchers are able to conduct larger experiments, is to design experiments to estimate those effects directly. At the most recent annual conference of the Bureau for Research and Economic Analysis of Development (the premier network for development economists) in May 2017, three of the eight papers presented described randomized controlled trials designed to assess the equilibrium impact of an intervention (Akram, Chowdhury, and Mobarak 2017; Andrabi et al. 2017; McKenzie and Puerto 2017). The typical design is a two-stage randomization procedure in which the treatment is randomly assigned at the market level in addition to the random assignment within a market. For example, the experiment of Crepon, Duflo, Gurgand, Rathelot, and Zamora (2013) varied the treatment density of a job placement assistance program in France within labor markets, in addition to random assignment of individuals within each market. The results show that placement assistance did benefit those assigned to receive it, but these effects were entirely undone by negative market-level impacts on untreated individuals. This result tempers the conclusion of a large literature on this type of intervention focusing on partial equilibrium effects, which tends to find that the program had significant positive effects (Card, Kluve, and Weber 2010). Muralidharan and Sundararaman (2015) adopt a similar design to evaluate a school voucher program in Andhra Pradesh, and in this case find no evidence of equilibrium effects coming into play.

A number of other experiments were designed to estimate just the full equilibrium effect, by conducting the randomization at the market level and focusing on market-level outcomes. Muralidharan, Niehaus, and Sukhtankar (2016) evaluate the rollout of a smart-card payments system for the National Rural Employment Guarantee Scheme, a workfare program in India. Randomization was conducted at the *mandal* (sub-district) level, allowing estimation of market-level effects across a large number of villages. The intervention increased take-up of the program, and the private sector wages increased in treatment *mandals* as a result. Several other papers estimate the impacts of transfer programs on village-level prices and wages (Cunha, De Giorgi, and Jayachandran 2011; Angelucci and De Giorgi 2009; Attanasio, Meghir, and Santiago 2011).

One potential challenge with the experimental identification of equilibrium effects is that it is not always obvious what the "market" is. For example, Akram,

Chowdhury, and Mobarak (2017) evaluate an intervention in rural Bangladesh that provided financial support for temporary migrants and find large effects on the migrants and their households. Implementation was randomized at the village level, as well as within villages, to examine spillover on nonparticipants (which is one type of general equilibrium effect), but the more obvious equilibrium effect in this case seems to be what happens to wages in cities when lots of migrants show up in a city. To address that, the randomization needs to be done at the level of the recipient city. This is conceptually feasible but a different exercise altogether (which this team plans to undertake in future research as the program scales).

One other form of general equilibrium effect receives less attention in the literature but can turn out to be relevant. When a particular intervention is scaled up, more people will be needed to implement it. This may lead to an increase in their wages or in difficulties hiring them, which should be accounted for in the cost–benefit analysis of the program at scale. For example, Duflo, Dupas, and Kremer (2017) exploit the result of their scholarship experiment to calculate the cost per year of making an extra year of secondary school free in Ghana. But once the government decides to implement free secondary schools in Ghana (as Sackey 2017 reports that they have just promised to do), the government will need to hire a large number of secondary schoolteachers. Given the short supply of college-educated workers, this may not be feasible or may be much more expensive than accounted for in the Duflo, Dupas, and Kremer (2017) calculations. The extent to which this is a problem in practice depends on the nature of the intervention and the context. Luckily, it seems researchers tend to be biased towards evaluating programs that do have a chance to be implementable at scale without a significant increase in costs.[1] A more general point is that any evaluation of benefits needs to be coupled with an understanding of the costs if it is to be useful as guidance for policy decisions. The costs will generally be different in the scaled-up version of the program than in the evaluation. Costs may in fact be lower once the program becomes routine—or higher, as in the Ghana case. Fortunately, a more accurate estimate of large-scale costs can often be estimated by collecting costs from versions of the programs that have been implemented at scale elsewhere.

Spillover Effects

Many treatments have spillovers on neighboring units, which implies that those units are not ideal control groups. Some spillovers are related to the technology: For example, intestinal worms are contagious, so if a child is dewormed, this will affect her neighbor. If many of the children in a school are dewormed, this will also affect neighboring schools (Miguel and Kremer 2004). An intervention targeted to some children in a school may also benefit others in the school who were in the control group—perhaps through peer effects or through adjustments in teaching within the school. Other channels of spillover are informational: when a new technology

[1] Banerjee, Duflo, and Kremer (2017) provide some tentative evidence suggesting that researchers are actually good at identifying such interventions before the experiment is conducted.

is introduced (like a long-lasting insecticide-treated bed-net), the first people who are exposed to it may not take it up or use it properly. As more people experience the product, their friends and neighbors will learn about it and moreover, this may have reinforcement effect as neighbors teach each other how to use it better. For example, Dupas (2014) evaluates the impact of free long-lasting insecticide-treated bed-net distribution in Kenya. She finds that when randomly selected households received a highly subsidized bed net in an initial distribution, their neighbors had a higher willingness to pay for a net one year later, suggesting they were learning about the technology.

Economists have long been mindful of the possibility of such spillovers, and even small experiments can be designed to investigate whether they are present. For example, Miguel and Kremer (2004) took advantage of the fact that the number of treatment schools was much higher in some areas than others (just by chance), to estimate the positive spillovers from taking the deworming medicine on those who did not themselves take it. Duflo and Saez (2003) adopt a two-step experimental design to measure information spillovers in retirement savings decisions. But not all spillovers are easy to detect in pilot experiments: in some cases, they may be highly nonlinear. For example, there may need to be enough people using a bed-net before substantial health externalities kick in: Tarozzi et al. (2014) conduct a randomized evaluation of the impact of bed-nets where the randomization was performed at the household level, and find no positive effect, but because very few households in each village received a bed-net, this does not tell us what would happen if they all got (and used) one. Cohen and Dupas (2010) show that calculations on the cost–benefit of free bed-net distribution are highly sensitive to assumptions made about nonlinear spillovers. This is potentially important given that standard models of social learning often embody important nonlinearities or "tipping points."

Political Reactions

Political reactions, including either resistance to or support for a program, may vary as programs scale up. Corrupt officials may be more likely to become interested in stealing from programs once they reach a certain size (Deaton 2010). For example, Kenya's national school-based deworming program, a scale-up based on the results of previous randomized controlled trials, began in 2009 but was halted for several years due to a corruption scandal. The funds for the program had been pooled with other funds destined for primary education spending, and allegations of misappropriation in those pooled funds caused donors to cut off education aid—including support for the deworming program. The program ultimately restarted in 2012 (Sandefur 2011; Evidence Action 2014).

Political resistance to or support for a program may build up when the program reaches a sufficient scale. Banerjee, Duflo, Imbert, Mathew, and Pande (2017) provide an example of political backlash leading to the demise of a promising program in the state of Bihar, India, to reduce corruption in a government workfare program. Even though the experiment was a pilot, it included almost 3,000 villages representing an overall population of 33 million people. The village officials and their immediate

superiors at the block- or district-level were dead set against the anticorruption intervention for the obvious reason that it threatened their rents. These officials were successful in lobbying the state government, and the intervention was cancelled, in part because a reduction in corruption was only demonstrated much later.[2]

This pilot of the anticorruption program was much larger than the typical proof-of-concept study, and as a result, the group it reached was large enough to have political influence. A smaller pilot might have had a less-difficult time, but this political counterreaction would have been missed. However, in other cases, pilots can be more vulnerable than scaled-up interventions: because they are subject to review, it is easy to shut them down.

Context Dependence

Evaluations are typically conducted in a few (carefully chosen) locations, with specific organizations. Would results extend in a different setting (even within the same country)? Would the results depend on some observed or unobserved characteristics of the location where the intervention was carried out?

Replication of experiments allows researchers to understand context dependence of programs. Systematic reviews, like those done by the Cochrane Collaboration for health care interventions, collect evidence from replications. Cochrane reviews have been compiled on topics such as water quality interventions (Clasen et al. 2015), mosquito nets (Lengeler 2004), and deworming of schoolchildren (Taylor-Robinson, Maayan, Soares-Weiser, Donegan, and Garner 2015). In economics, the International Initiative for Impact Evaluation maintains a database of systematic reviews of impact evaluations in developing countries that contains more than 300 studies (International Initiative for Impact Evaluation 2017). Several recent studies and journal volumes compile the results from multiple interventions in the same publication. For example, the January 2015 issue of the *American Economic Journal: Applied Economics* was devoted to six experimental studies of microfinance. Although these studies were not conducted in coordination, the overall conclusions are quite consistent across studies: the interventions showed modest increases in business activity but very little evidence of increases in consumption (Banerjee, Karlan, and Zinman 2015). The development of the American Economic Association's registry of randomized trials and public archiving of data, and the greater popularity of systematic meta-analysis methods within economics, should allow similar analyses across many more research questions.[3]

[2] There was, however, an interesting postscript: The results—which came out after the pilot was cancelled in Bihar—indicated a significant decline in rent-seeking and the wealth of public program officials. The anticorruption program was then extended to the same workfare program in all of India (with an explicit reference to the experimental results), and there are discussions to extend it to other government transfer programs.

[3] McEwan (2015) is another example of meta-analysis. He analyzes the results of 77 randomized controlled trials of school-based interventions in developing countries that examine impacts on child learning. While there is some degree of heterogeneity across studies, he is able to classify types of interventions that are consistently most effective based on his random-effects model.

However, to aggregate effects across studies, one has to start from some assumption about the potential distribution of treatment effects (Banerjee, Chassang, and Snowberg 2017). In the economics literature, this is often done without a formal analytical framework, which can lead to misleading results. For example, Pritchett and Sandefur (2015) argue that context-dependence is potentially very important, and that the magnitude of differences in treatment effects across contexts may be larger than the magnitude of the bias generated from program evaluation using retrospective data. They illustrate their point with data from the six randomized controlled trials of microcredit mentioned above. However, as pointed out by Meager (2016), Pritchett and Sandefur's measure of dispersion grossly overstates heterogeneity by conflating sampling variation with true underlying heterogeneity. Meager applies to the same data a Bayesian hierarchical model popularized by Rubin (1981), which assumes that (true) treatment effects in each site are drawn randomly from a normal distribution, and then estimated with error, and finds remarkably homogenous results for the mean treatment effect.

However, once we admit the need for a prior for aggregating results, there is no reason to stick to purely statistical approaches. An alternative is to use the existing evidence to build a theory, which tries to account for why some experiments succeed and others fail—rather than just tallying all the experiments and letting the failures cancel out the successes. The theory can then offer predictions that could be tested in future experiments, or which can feed into the design of a scaled-up intervention. For example, Kremer and Glennerster (2011) consider a range of randomized controlled trials on how price sensitivity affects take-up of preventive health products. They propose a number of alternative theories featuring liquidity constraints, lack of information, nonmonetary costs, or behavioral biases (such as present bias and limited attention). Dupas and Miguel (2017) provide an excellent summary of the evidence from randomized controlled trials on this point and argue that the subsequent evidence supports some aspects of the Kremer–Glennerster framework and rejects others. The point here is that many of those subsequent experiments were designed precisely with the Kremer–Glennerster framework in mind—effectively testing their conjectures—which makes them much more informative.

Randomization or Site-Selection Bias

Organizations or individuals who agree to participate in an early experiment may be different from the rest of the population, which Heckman (1992) calls randomization bias. There are three different possible sources for this concern.

First, organizations (and even individuals within governments) who agree to participate in randomized controlled trials are often exceptional. Glennerster (2017) lists the characteristics of a good partner to work with for a randomized controlled trial, and many organizations in developing countries do not meet the criteria. For example, the organization must be able to organize and implement the randomized implementation, providing relatively uniform implementation in the treatment group while not contaminating the control group. Senior staff must be

open to the possibility of the program not working and be willing to have these results publicized. Even within government, willing partners are often particularly competent and motivated bureaucrats. Even when an intervention is not "gold-plated," organizations of individuals with these capabilities may find larger effect sizes than a large-scale program run by a less-stellar organization.[4] This is different from the general equilibrium point made above—even when the personnel to carry out the intervention at scale exists, the key constraint may be that of management capacity, and the difficulty of implementing changes at scale.

Second, a well-understood problem arises when individuals select into treatment. If treatment effects are heterogeneous across these groups, and those who are more likely to benefit are also more likely to be treated, then the estimated effect from the randomized controlled trial applies to compliers (those that respond to treatment), and may not apply to a broader population (Imbens and Angrist 1994). However, randomized controlled trials can be designed to enhance the external validity of experiments when respondents select themselves into treatment (for the theory, see Chassang, Padró i Miquel, and Snowberg 2012; for an application, see Berry, Fischer, and Guiteras 2015).

Third, site-selection bias arises because an organization chooses a location or a subgroup where effects are particularly large. This choice could be for legitimate reasons: nongovernmental organizations have limited resources and will try to work where they think their impact is the greatest, so they go to those areas first. In addition, both the organizations and the researchers, knowing that they are subject to an evaluation, have incentives to choose a site where the program is more likely to work. Organizations who take the trouble to participate in a randomized controlled trial would rather demonstrate success. Furthermore, if researchers anticipate that a study finding significant results is more likely to be published, they may design their studies accordingly. An illustrative case is that of Banerjee, Barnhardt, and Duflo (2015), who find no impact on anemia of free iron-fortified salt, in contrast with previous randomized controlled trials which led to the approval of the product for general marketing. And one reason is that the previous studies targeted adolescent women—and in fact Banerjee, Barnhardt, and Duflo (2015) find substantial treatment effects for that group but not for the average person. Yet fortified salt was approved for sales and distribution to the general population based on the group-specific results.

Several recent papers examine issues of randomization bias across large numbers of studies. Vivalt (2016) compiles data from over 400 randomized controlled trials and examines the relationship between effect size and study characteristics. Studies evaluating interventions run by nongovernment organizations or by researchers tend to find higher effects than randomized controlled trials run with governments, as do studies with smaller sample sizes. Allcott (2015) presents

[4] Allcott (2015) compares microfinance institutions that have partnered in recent randomized controlled trials with a global database of microfinance institutions and finds that partner institutions are older, larger, and have portfolios with lower default risk compared with nonpartner institutions.

the results of 111 randomized controlled trials of the Opower program in which households are presented with information on energy conservation and energy consumption of neighbors. He finds that the first ten evaluations of the intervention show larger effects on energy conservation than the subsequent evaluations and argues that this finding is attributable to differences in both partner utilities and study populations. Blair, Iyengar, and Shapiro (2013) examine the distribution of randomized controlled trials across countries and find that such trials are disproportionally conducted in countries with democratic governments.

Piloting Bias/Implementation Challenges

A large-scale program will inevitably be run by a large-scale bureaucracy. The intense monitoring that is possible in a pilot may no longer be feasible when that happens, or may require a special effort. For example, school reform often requires buy-in from teachers and school principals to be effective. The Coalition for Evidence-Based Policy (2013) reviewed 90 evaluations of US educational interventions commissioned by the Institute of Educational Studies, the research arm of the US Department of Education. They found that lack of implementation by the teachers was a major constraint and one important reason why 79 of 90 these interventions did not have positive effects. Interestingly, these interventions were themselves often quite small scale, despite being scale-ups of other even smaller studies.

Studies rarely document implementation challenges in great detail, but there are some examples. Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2015) replicate an intervention in Kenya first evaluated in Duflo, Dupas, and Kremer (2011, 2015), in which a nongovernment organization gave grants to primary school parent–teacher associations to hire extra teachers in order to reduce class sizes. The original Duflo, Dupas, and Kremer (2011, 2015) intervention resulted in significant increases in test scores. Bold et al. (2015) evaluate two versions of the program: one run by a nongovernment organization, which produced very similar results to the Duflo, Dupas, and Kremer (2011, 2015) evaluation, and a government-run version, which did not produce significant gains. Analysis of process data finds that government implementation was substantially weaker: the government was less successful in hiring teachers, monitored the teachers less closely, and was more likely to delay salary payments. The authors also suggest that political reactions—particularly the unionizing of the government contract teachers—could have also dampened the effects of the government-led implementation.

A number of studies have found differences between implementation by nongovernment organizations and governments. Barrera-Osorio and Linden (2009) evaluate a program in Colombia in which computers were integrated into the school language curriculum. In contrast with a previous intervention led by a nongovernment organization in India (Banerjee, Cole, Duflo, and Linden 2007), the authors find negligible effects of the program on learning, which they attribute to the failure of the teachers. Banerjee, Duflo, and Glennerster (2008) report on an experiment where incentives were provided for verified attendance in government

health clinics in India. Although a similar incentive scheme had previously been proven effective when implemented in education centers run by a nongovernment organization in the same area (Duflo, Hanna, and Ryan 2012), there were no long-term effects on attendance in government health centers due to staff and supervisors exploiting loopholes in the verification system. Banerjee, Chattopadhyay, Duflo, Keniston, and Singh (2014), working with the police leadership in Rajasthan, India, to improve the attitudes of the police towards the public, find that the reforms that required the collaboration of station heads were never implemented.

In an interesting counterexample, Banerjee, Hanna, Kyle, Olken, and Sumarto (2016) study the distribution of identity cards entitling families to claim rice subsidies in Indonesia. In the pilot, the Indonesian government was meant to distribute cards containing information on the rice subsidy program to beneficiary households, but only 30 percent of targeted households received these cards. When the program was scaled up to the whole country, the mechanism for sending cards was changed and almost everybody did finally get a card. In this case, the government was less effective at running a pilot program and more effective with full implementation. This dynamic may be more general than one might at first expect: pilots face their own challenge because they impose new ad hoc procedures on top of an existing system. Once a bureaucracy takes over and puts a routine in place, implementation can become more systematic.

As the discussion in this section has emphasized, the issue of how to travel from evidence at proof-of-concept level to a scaled-up version cannot be settled in the abstract. The issue of context-dependence needs to be addressed through replications, ideally guided by theory. General equilibrium and spillover effects can be addressed by incorporating estimation of these effects into study designs, or by conducting large-scale experiments where the equilibrium plays out. Randomization and piloting bias can be addressed by trying out the programs on a sufficient scale with the government that will eventually implement it, documenting success and failure, and moving from there.

In the next section, we illustrate how these issues play out in practice by describing the long journey from the original concept of a specific teaching intervention in India, through its multiple variants, to the eventual design and evaluation of two "large-scale" successful incarnations implemented in government schools that are now in the process of being scaled up in other government systems.

A Successful Scale-up: Teaching at the Right Level

In India, as in many developing countries, teachers are expected to teach a demanding curriculum, regardless of the level of preparation of the children. As a result, children who get lost in early grades may never catch up (Muralidharan 2017). In response, Pratham, an Indian nongovernmental organization, designed a deceptively simple approach, which has come to be called "teaching at the right level" (TaRL). Pratham credits literacy expert Abul Khair Jalaluddin for developing

the first incarnation of the pedagogy (Banerji, Chavan, and Rane 2004). The basic idea is to group children, for some period of the day or part of the school year, not according to their age, but according to what they know—for example, by splitting the class, organizing supplemental sessions, or reorganizing children by level—and match the teaching to the level of the students.

From Bombay Slums to 33 Million Children

The partnership between researchers and Pratham started with a "proof of concept" randomized controlled trial of Pratham's *Balsakhi* Program in the cities of Vadodara and Mumbai, conducted in 2001–2004 (Banerjee et al. 2007). In this program, third- and fourth-grade students identified as "lagging behind" by their teachers were removed from class for two hours per day, during which they were taught remedial language and math skills by paid community members (*balsakhis*) hired and trained by Pratham. Their learning levels (measured by first- and second-grade-level tests of basic math and literacy) increased by 0.28 standard deviations.

Pratham next took this approach from the relatively prosperous urban centers in West India into rural areas, and in particular into rural areas of Northern India. By 2004, Pratham worked in 30 cities and nine rural districts (Banerji, Chavan, and Rane 2004). As Pratham increased the scale of its program, the key principle of teaching children at the appropriate level remained, but one core feature of its model changed for the sake of financial viability: they were forced to rely largely on volunteers rather than paid teachers. These volunteers worked outside the school running their own learning-improvement classes and were much less closely supervised after the initial two-week training. To facilitate this change, the pedagogy became more structured and more formal, with an emphasis on frequent testing. Whether the intervention would continue to work well with a new programmatic design, organizational change, and new contexts was an open question. A new randomized evaluation was therefore launched to test the volunteer-based model in the much more challenging context of rural North India.

This second randomized controlled trial was conducted in rural Jaunpur district of Uttar Pradesh in 2005–2006: this was a test of the volunteer-led, out-of-school model Pratham called "Learning to Read." The results were very positive: after accounting for the fraction of students who attended, treatment-on-the-treated estimates show that attending the classes made children who could read nothing at baseline 60 percent more likely to progress to letters at endline. For children who could read letters at baseline, the classes resulted in a 26 percent higher likelihood of reading a story, the highest level on the test, at endline (Banerjee, Banerji, Duflo, Glennerster, and Khemani 2010).

This second study established that the pedagogical idea behind the *balsakhi* program could survive the change in context and program design, but it also revealed new challenges. There was substantial attrition among the volunteers, and many classes ended prematurely. Also, because the program targeted children outside of school, take-up was far from universal. Only 17 percent of eligible

students were treated. Most concerning, the treated students did not come disproportionately from the bottom end of the distribution—those who were unable to recognize letters or numbers, and who needed it the most.

Nevertheless, in 2007, building on the success of the Learning to Read intervention, Pratham rolled out its flagship "Read India" Program. Within two years, the program reached over 33 million children. To reach all of the children who needed remedial education, Pratham started collaborating with state governments in running the program. But the efficacy of the government's implementation of the program was again an open question. In the remainder of this section, we present the results of the series of experiments aimed to develop a scalable policy in government schools based on the Pratham methodology.

A First Attempt to Scale-Up with Government

Starting in 2008, Pratham and the Abdul Latif Jameel Poverty Action Lab, commonly known as J-PAL, embarked on a series of new evaluations to test Pratham's approach when integrated with the government school system. Two randomized controlled trials were conducted in the states of Bihar and Uttarakhand over the two school years of 2008–2009 and 2009–2010. Although the evaluation covered only a few hundred schools, it was embedded in a full scale-up effort: as of June 2009, the Read India program was being run in approximately 40,000 schools in Bihar and 12,000 schools in Uttarakhand, representing the majority of schools in each state (Kapur and Icaza 2010).

In the first intervention (evaluated only in Bihar during June 2008), remedial instruction was provided during a one-month summer camp, run in school buildings by government schoolteachers. Pratham provided materials and training for these teachers and also trained volunteers who supported teachers in the classroom. The government schoolteachers were paid extra by the government for their service over the summer period.

The other three interventions were conducted during the school year. The first model (evaluated only in Bihar) involved the distribution of Pratham materials with no additional training or support. The second variant of the intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff. Teachers were trained to improve teaching at all levels through better targeting and more engaging instruction. The third and most-intensive intervention included materials, training, and volunteer support. The volunteer part of the materials-training-volunteers intervention in Bihar was a replication of the successful Learning-to-Read model evaluated in Jaunpur, in which the volunteers conducted out-of-school learning camps that focused on remedial instruction for students directed to them by teachers. As part of the materials-training-volunteers intervention in Uttarakhand, however, volunteers worked in schools and were meant to support the teachers. In both states, about 40 villages were randomly assigned to each treatment group.

The main outcome measures in the Bihar and Uttarakhand evaluations, as with the others presented later in this section, are performance on simple language and

math tests developed by the ASER Centre, Pratham's research arm. In language, children are classified based on whether they can recognize letters, read words, read a paragraph, or read a short story. In math, the levels include single-digit and double-digit number recognition, double-digit subtraction, and division of a double-digit number by a single digit. In the results that follow, we assign an integer score between zero and four based on the highest level the child can perform.

To complement the randomized controlled trial, we collected extensive process data and partnered with political scientists who, through interviews, collected details of the relationship between Pratham and the government.[5]

The language and math results of the evaluations in Bihar and Uttarakhand (presented in Table 1) were striking and mostly disappointing. The materials-alone and materials-plus-training interventions had no effect in either Bihar or Uttarakhand. The materials-training-volunteers treatment in Uttarakhand had no detectible impact either. However, in the materials-training-volunteers results in Bihar, we found a significant impact on reading and math scores, quite comparable to the earlier Jaunpur results. Since the materials-plus-training intervention seemed to make no difference, we interpret this as further evidence that, like in Jaunpur, Pratham's pedagogical approach also worked in this new context when implemented by volunteers outside school hours. However, when the volunteers were made part of the in-school team, as in Uttarakhand, they either became absorbed as regular teachers, teaching the curriculum rather than implementing Pratham's pedagogy, or did not show up at all. The failure of schools to utilize the volunteers as intended may be why the Uttarakhand intervention did not work.

At this point, one might have been tempted to assume that the key distinction is between government teachers and private volunteers as implementers (along the lines of Bold et al. 2015). However, this interpretation is belied by the Bihar summer camp results, which show significant gains in language and math despite being implemented by the government schoolteachers. Based on the fraction of children who attended the summer camp, the treatment-on-the-treated results show that the camp improved reading scores by about 0.5 levels in just a few weeks. This finding suggests the possibility that government teachers were in fact able to deliver remedial education if they did focus on it, but this did not happen during the school year.

Some process data and the qualitative information bolster this interpretation. Table 2 (panels A and B) shows selected process measures across the two experiments. The situations were very different in the two states (Kapur and Icaza 2010). In Bihar, Pratham had an excellent relationship with the educational bureaucracy, from the top rungs down to district- and block-level administrators. As a result, the basic inputs of the program were effectively delivered: over 80 percent of the

[5] Banerjee, Banerji et al. (2016) provide more details on the evaluation design and the results of these two experiments as well as the two further experiments described in the next subsection. Kapur and Icaza (2010) provide a detailed account of the working of the partnership between Pratham and the government at various levels in Bihar and Uttarakhand. Sharma and Deshpande (2010) present a qualitative study based on interviews with parents, teachers, and immediate supervisors of the teachers.

Table 1
Language and Math Results—Bihar and Uttarakhand

	Test Score (0–4)	
	Language	*Math*
A. Bihar—Summer Camp		
Treatment	0.12**	0.085*
	(0.059)	(0.050)
Control group mean	2.2	2.1
Observations	2,839	2,838
B. Bihar—School Year		
Materials	0.027	0.051
	(0.061)	(0.051)
Materials, Training	0.064	0.017
	(0.059)	(0.049)
Materials, Training, Volunteer Support	0.20***	0.13***
	(0.054)	(0.046)
Control group mean	1.8	1.8
Observations	6,490	6,490
C. Uttarakhand		
Materials, Training	0.030	0.038
	(0.053)	(0.042)
Materials, Training, Volunteer Support	–0.012	0.0091
	(0.044)	(0.043)
Control group mean	2.2	2.0
Observations	5,645	5,646

Note: Pratham and the Abdul Latif Jameel Poverty Action Lab (J-PAL) conducted randomized controlled trials testing the Pratham pedagogical approach when integrated with the government school system in the states of Bihar and Uttarakhand. In the first intervention (Panel A), remedial instruction was provided during a one-month summer camp run in school buildings by government schoolteachers. Pratham provided materials and training for these teachers and also trained volunteers who supported teachers in the classroom. The other three interventions (Panels B and C) were conducted during the school year: The first model (evaluated only in Bihar) involved the distribution of Pratham materials with no additional training or support. The second intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff. The third and most-intensive intervention included materials, training, and volunteer support. In Bihar, the volunteers conducted out-of-school learning camps (during the school year) that focused on remedial instruction for students directed to them by teachers. As part of the materials-training-volunteers intervention in Uttarakhand, however, volunteers worked in schools and were meant to support the teachers. Standard errors in parentheses (clustered at the level of randomization). Test scores are computed on an integer scale from 0 (nothing) to 4 (can read a story) in language and 0 (nothing) to 4 (can perform division) in math. Regressions control for baseline scores as well as gender age, and grade at baseline.
*, **, and *** mean significance at the 10, 5, and 1 percent levels, respectively.

Table 2
Selected Process Results

	Percent of schools (# of schools in parentheses)		
	Teachers trained	*Pratham materials used*	*Classes grouped by ability*
A. Bihar—School Year			
Control	1.4	0.8	0.0
	(63)	(59)	(60)
Materials	5.6	33.6	1.6
	(64)	(63)	(63)
Materials, Training	84.4	62.5	3.8
	(66)	(64)	(65)
Materials, Training, Volunteer Support	84.7	69.2	0.0
	(68)	(65)	(65)
B. Uttarakhand			
Control	18.9	3.8	14.1
	(41)	(39)	(39)
Materials, Training	29.4	26.3	10.0
	(40)	(40)	(40)
Materials, Training, Volunteer Support	53.8	38.5	5.1
	(39)	(39)	(39)
C. Haryana			
Control	0.5	0.5	0.0
	(200)	(199)	(199)
Teaching at the Right Level (During TaRL classes)	94.7	81.0	91.3
	(126)	(126)	(126)
Teaching at the Right Level (Other times)	94.0	1.3	2.0
	(155)	(149)	(149)
D. Uttar Pradesh			
Control		0.0	
		(108)	
Materials		30.7	
		(111)	
Four 10-Day Camps	89.9	90.6	79.4
	(122)	(122)	(122)
Two 20-Day Camps	87.8	84.2	83.5
	(120)	(120)	(120)

Note: The Bihar school-year and Uttarakhand evaluations consisted of three interventions. The first model (evaluated only in Bihar) involved the distribution of Pratham materials with no additional training or support. The second intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff. The third and most-intensive intervention included materials, training, and volunteer support. In the Haryana intervention, efforts were made to promote organizational buy-in, including the creation of a system of academic leaders within government to guide and supervise the teachers as they implemented the Pratham methodology; the program was implemented during a dedicated hour of the school day; and all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels. The Uttar Pradesh interventions used the in-school "learning camps" model, with learning camps administered primarily by Pratham volunteers and staff during school hours when regular teaching was temporarily suspended. When a school was observed multiple times, the average is used for that school.

teachers were trained, they received the material, and they used the materials more than half the time. In Uttarakhand, key state personnel changed just before the evaluation period and several times afterwards. There was infighting within the educational bureaucracy, and strikes by teachers and their supervisors (unrelated to the program). The local Pratham staff were demoralized and turned over rapidly. As a result, only between 29 and 54 percent of teachers got trained (for only three days each), and only one-third of the schools used the materials, which they received very late. In many cases, there was either no volunteer or no teacher in the school during the monitoring visits.

The process data also show that the key component of Pratham's approach, the focus on teaching at the children's level, were generally not implemented in schools in either state. One consistent lesson of the earlier studies is that the pedagogy worked when children grouped in a way that the teaching could be targeted to the deficiencies in their training. This happened systematically in the volunteer classes, and this also happened in the Bihar summer camps because their express purpose was to focus on remedial skills. But in regular classes in Bihar, for example, only between 0 and 4 percent of the classes were observed to be grouped by levels.

Thus, the challenge for Pratham was how to get government teachers to not only use materials and deliver the pedagogy, but also how to incorporate the targeted teaching aspect of the model into the regular school day. As we see from Bihar, independent training by Pratham by itself was insufficient to get teachers to do this, even with consistent support by the bureaucracy. The summer camp in Bihar, however, produced a large effect. Therefore, it is possible for governments to "teach at the right level." Why don't they do so during the regular school day?

In *Poor Economics*, Banerjee and Duflo (2011) discuss this resistance and point out that Teaching at the Right Level is not even being implemented in private schools, which are subject to a high level of competition and are certainly not lacking in incentives, despite the fact that most children in those schools are also not at grade level. They propose the hypothesis that teachers and parents must put more weight on covering the grade-level curriculum than on making sure that everyone has strong basic skills. Similarly, the qualitative interviews conducted in the Read India scale-up experiments revealed that teachers believed the methods proposed by Pratham were effective and materials were interesting, but they did not think that adopting them was a part of their core responsibility. Paraphrasing the teachers they interviewed in Bihar, Sharma and Deshpande (2010) write: "[T]he materials are good in terms of language and content. The language is simple and the content is relevant. … However, teaching with these materials require patience and time. So they do not use them regularly as they also have to complete the syllabus."

If this hypothesis is correct, it suggests two main strategies: either convince the teachers to take Teaching at the Right Level more seriously by working with their superiors to build it into their mission; or cut out the teachers altogether and implement a volunteer-style intervention, but do it in the school during school

hours, so as to capture the entire class rather than just those who opt to show up for special after-school or summer classes. These ideas guided the design of the next two interventions.

Getting Teachers to Take the Intervention Seriously

In 2012–2013, Pratham, in partnership with the Haryana State Department of Education, adopted new strategies to embed the Teaching at the Right Level approach more strongly into the core of teaching/learning in primary schools; in particular, they were interested in how to get teachers to view it as a "core responsibility."

Several methods were used to promote organizational buy-in. First, all efforts were made to emphasize that the program was fully supported and implemented by the government of Haryana, rather than an external entity. In the earlier experiment in Bihar and Uttarakhand, despite the fact that this was a government initiative, teachers did not perceive it as such, in part because they rarely got that message from their immediate supervisors and the responsibility of monitoring the teachers was left to Pratham staff. In Haryana, a system of academic leaders within the government was created to guide and supervise teachers as they implemented the Pratham methodology. As part of the interventions, Pratham gave four days of training and field practice to "Associate Block Resources Coordinators," who were then placed in groups of three in actual schools for a period of 15–20 days to carry out daily classes and field-test the Pratham methodology of grouping by level and providing level-appropriate instruction. Once the practice period was over, these Coordinators, assisted by Pratham staff, in turn trained the teachers that were in their jurisdiction. Second, the program was implemented during a dedicated hour during the school day. Beginning in the 2011–2012 school year, the government of Haryana mandated that all schools add an extra hour of instruction to the school day. In regular schools, the normal school day was just longer. Within Teaching at the Right Level schools, the extra hour was to be used for class reorganization and teaching remedial Hindi classes using the Pratham curriculum. This change sent a signal that the intervention was government-mandated, broke the status quo inertia of routinely following the curriculum, and made it easier to observe compliance. Third, during the extra hour, in Teaching at the Right Level schools, all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels, as determined by a baseline assessment done by teachers and the coordinators. Once classes were restructured into these level-based groups, teachers were allocated to the groups for instruction. This removed teacher discretion on whether to group children by achievement.

This new version of the program was evaluated in the school year 2012–2013 in 400 schools, out of which 200 received the program. The results were this time positive, as shown in Table 3: Hindi test scores increased by 0.2 levels. This intervention did not target math.

Because the objective of this study was to develop a model that could be adopted at scale, we also incorporated extensive process monitoring into our study design,

Table 3
Language and Math Results—Haryana and Pradesh

	Test Score (0–4)	
	Language	*Math*
A. Haryana		
Teaching at the Right Level	0.20***	–0.0069
	(0.023)	(0.019)
Control group mean	2.4	2.2
Observations	11,963	11,962
B. Uttar Pradesh		
Materials	0.045	0.053**
	(0.030)	(0.027)
Four 10-Day Camps	0.95***	0.81***
	(0.030)	(0.028)
Two 20-Day Camps	0.82***	0.73***
	(0.031)	(0.029)
Control group mean	1.5	1.7
Observations	17,254	17,265

Note: In the Haryana intervention, efforts were made to promote organizational buy-in, including the creation of a system of academic leaders within government to guide and supervise the teachers as they implemented the Pratham methodology; the program was implemented during a dedicated hour of the school day; and all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels. The Uttar Pradesh interventions used the in-school "learning camps" model, with learning camps administered primarily by Pratham volunteers and staff during school hours when regular teaching was temporarily suspended. Standard errors are in parentheses (clustered at the level of randomization). Test scores are computed on an integer scale from 0 (nothing) to 4 (can read a story in language, and 0 (nothing) to 4 (can perform division) in math. Regressions control for baseline test scores, as well as gender, age, and grade at baseline.
*, **, and *** mean significance at the 10, 5, and 1 percent levels, respectively.

including regular surprise visits to the schools. The third panel of Table 2 shows that about 95 percent of teachers in the treatment group attended training, compared with virtually no teachers in the control group. Most importantly, grouping by ability was also successful in Haryana, where it had largely failed in Bihar and Uttarakhand: over 90 percent of schools were grouped by learning levels during Teaching at the Right Level classes. In addition, teachers in Haryana used Pratham materials in 81 percent of Teaching at the Right Level classes, whereas much lower rates were observed in Bihar and Uttarakhand. Interviews with teachers, headmasters, and department administration suggested that the monitoring and mentoring role played by Associate Block Resource Coordinators was critical. Indeed, 80 percent of schools reported a visit from a Coordinator in the previous 30 days. Of those who

reported a visit, 75 percent said that the Coordinator spent over an hour in the school, and 95 percent said that the Coordinator observed a class in progress.

Using the Schools, But Not the Teachers: In-School Learning Camps

In areas where the teaching culture is very weak, it may be too difficult or costly to involve the teachers in this alternative pedagogy. Instead, it may make sense to use an outside team, which can sidestep the teachers but still take advantage of the school infrastructure and the fact that the children are already present at school. The risk in going down this path, as we had seen in Uttarakhand before, was that the volunteers would be absorbed by the system and not implement the desired pedogogy.

To address this issue, Pratham, with the permission of the district administration, developed the in-school "Learning Camps" model. Learning Camps are intensive bursts of teaching/learning activity using the Pratham methodology and administered primarily by Pratham volunteers and staff during school hours when regular teaching is temporarily suspended. These camps were confined to periods of 10 or 20 days each (and a total of 50 days a year). In that sense, they were more similar to the original volunteer Learning-to-Read model (where volunteers ran "sessions" of 2–3 months) than to previous in-school experiences, except that they were within school premises during school hours. On "camp" days, children from grades 3–5 were grouped according to their ability level and taught Hindi and math for about 1.5 hours each by Pratham staff and Pratham-trained local village volunteers.

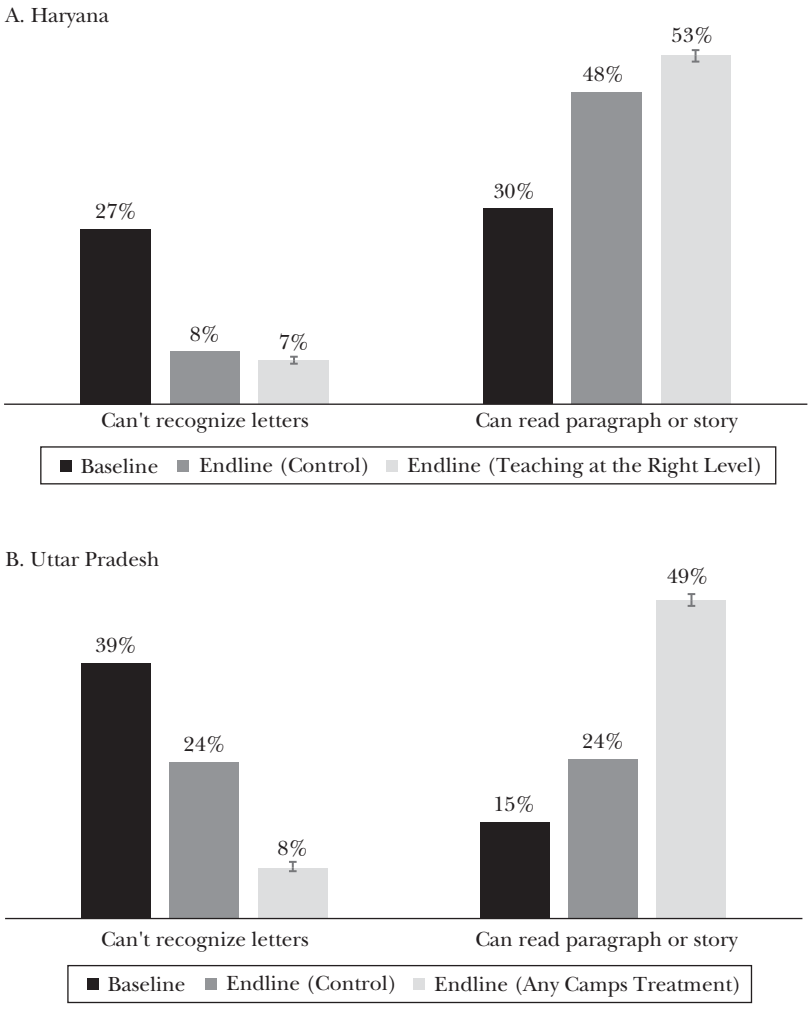
The model was tested in a randomized evaluation in Uttar Pradesh in the year 2013–2014: a sample of schools was selected and randomly divided into two camp treatment groups, a control group, and a materials-only intervention, with approximately 120 schools in each group. The learning camp intervention groups varied the length of the camp rounds, with one group receiving four 10-day rounds of camp, and the second receiving two 20-day rounds. Each intervention included an additional 10-day camp during the summer.

The two interventions had similar impacts, with test score gains of 0.7 to 1.0 levels, on average (as shown in Table 3).

It is useful to pause and consider how large these effects are. Figure 1 summarizes the results in Haryana and Uttar Pradesh. At baseline, 27 percent of children in Haryana could not even recognize letters, and 30 percent could read a paragraph or story (since the studies are randomized, control group and treatment group students are similar, so we present the pooled data for the baseline statistics). In Uttar Pradesh, 39 percent of the children could not recognize letters, and only 15 percent could read a paragraph or story. The difference between the two states was very large. At endline, there was little progress in the control group in Uttar Pradesh: 24 percent of children could still not recognize letters, and only 24 percent could read a paragraph or a story. But in the treatment group, only 8 percent could not recognize letters, and 49 percent could read a paragraph or a story. Thanks to these 50 days of intervention, they had fully caught up to the control group in Haryana

Figure 1

Distribution of Student Competency in Language: Baseline and Endline, by Treatment Status



Note: In the Haryana intervention, efforts were made to promote organizational buy-in, including the creation of a system of academic leaders within government to guide and supervise the teachers as they implemented the Pratham methodology; the program was implemented during a dedicated hour of the school day; and all children in grades 3–5 were reassigned to achievement-based groups and physically moved from their grade-based classrooms to classrooms based on levels. The Uttar Pradesh interventions used the in-school "learning camps" model, with learning camps administered primarily by Pratham volunteers and staff during school hours when regular teaching was temporarily suspended. Whiskers represent the 95 percent confidence interval between intervention and control groups.

(where, at endline, 48 percent could read a paragraph or story and 8 percent could not recognize letters), and had almost reached the level of the treated children in Haryana (where 53 percent of the treatment children could read a story). This reflects in part the abysmal performance of the school system in Uttar Pradesh. But the fact that the children actually reach the Haryana level in Uttar Pradesh also demonstrates the relative ease with which apparently daunting learning gaps can be closed.

As with the other evaluations, a systematic process-monitoring survey collected data on attendance, evidence of learning by "grouping," activities during "camp" sessions, teaching practices of volunteers, involvement of schoolteachers, and their perception of "camp" activities. There was strong adherence to key program components in Uttar Pradesh (Table 2, panel D). During camp days, use of Pratham materials was observed in over 80 percent of classes in both the 10-day and 20-day camp interventions. Critically, about 80 percent of classes in both treatments were observed to be grouped by achievement.

It took five randomized control trials and several years to traverse the distance from a concept to a policy that actually could be successful on a large scale. Today, the teacher-led "Haryana" model has been implemented in 107,921 schools across 13 states of India, reaching almost 5 million children. The in-school volunteer led model has been implemented in 4,210 schools across India, reaching over 200,000 children.

General Lessons about Scaling Up

Of the potential scale-up issues we identified, which ones turned out to be relevant in the Teaching at the Right Level example, and beyond that, what should be taken into consideration when designing an experiment with the view of ultimate scale up?

Equilibrium effects were not really a threat in this context, despite the size of the scale up in which the evaluations were embedded, since our outcome of interest was human capital, where there is no strategic interdependence. We did not explicitly study *spillovers* (although some could have occurred between teachers).

The interventions were repeatedly stress-tested for *context dependence* by moving the program from urban India to rural Uttar Pradesh, and then to Bihar, Uttarakhand, Haryana, and back to Uttar Pradesh again. This shows that the pedagogy that Pratham developed can improve basic learning levels in both reading and math across a variety of contexts. Moreover, there is supporting evidence from Ghana, where a successful replication of the Teaching at the Right Level approach was organized with the government (Innovations for Poverty Action 2015), and in Kenya, where students performed better when grouped by ability (Duflo, Dupas, and Kremer 2011). The results both in India, alongside results from similar tests worldwide, made it clear that many children clearly needed remedial education. In terms of understanding the magnitude of the need for remedial education, it is striking

that the *intention-to-treat* effect of the camps program in Uttar Pradesh (estimated over all children in the schools) are as high as the *treatment-on-the-treated* results were in the early Learning to Read evaluation in Jaunpur (estimated only for those that attended after-school classes) (Banerjee et al. 2010). This finding suggest that the high early results were not driven by a subpopulation with high marginal returns in the original experiment.

Although political issues did arise in Uttarakhand, they were more due to turnover and infighting than to issues with Pratham, and there were no direct adverse *political reactions* to the program in its scaled-up version. However, such resistance could arise elsewhere. An attempt to pilot the project in the state of Tamil Nadu was not successful after the government officials displayed strong resistance. The backstory here is that Pratham has become such an important large player in the India educational scene that it cannot be seen as just another partner organization. In Tamil Nadu, Pratham was viewed as the group that had exposed the less-than-stellar performance of the state-run schools. Also, the Tamil Nadu government had their own pedagogical approach called "Activity Based Learning," which it was not keen to subject to scrutiny.

Although most of the attention on the challenge of scalability in the recent literature has been on the equilibrium effects and context dependence, it appears these issues were not particularly relevant here. The key obstacle to Teaching at the Right Level was the difficulty of implementing at scale. The first successes were clearly affected by a combination of *randomization bias* and *implementation challenges* when moving to scale. Pratham was one of the first organizations to partner with researchers to evaluate its programs (before J-PAL even existed), and may be rare in its combination of scale and purpose. It is conceivable that moving from Pratham to *any* other partner, not just the government, would have been difficult. Even within Pratham it was harder to find a good and enthusiastic team in Uttarakhand than in Bihar (Kapur and Icaza 2010). The fundamental challenge was to integrate the core concept of the program in the schools' day-to-day workflow, and this relied on organizational innovations beyond the basic concept of Teaching at the Right Level. In particular, achieving the alignment between pedagogy and initial learning levels required an explicit organizational effort to ensure that children were assessed, grouped, and actually taught at the right level. This did not occur automatically within the existing government school system but was achieved by articulating a narrow set of program goals, ensuring there was specific time for the program, and properly supervising implementation.

One way to interpret the series of Teaching at the Right Level studies is as a process of persuasion at scale: the experimental approach played not only an evaluation role but also an instrumental role in fostering acceptance of the policy by the government. In other words, we can see this effort as trying to answer the question: "How do you get a bureaucracy to make a common-sense change that has a very strong chance of being beneficial—like not totally ignoring students who have fallen behind and instead offering them a path to catching up?" From that perspective, the experimental approach is a little like opening a jammed door with a pry-bar. First

you stick the bar in a little crack and get a little traction. Then you move to another location and get a little more traction. When you've got a little more purchase, you can jam in a bigger pry-bar and really tug hard. From this perspective, choosing where to pry, and finding organizations willing to experiment, and choosing other places to pry, and then finding government partners willing to participate, is all a way of prying open the door. At some point, the leverage is great enough that you can throw the door open. Sequential experimentation becomes a political economy tool for getting momentum for policy change.

More generally, what should practitioners and researchers keep in mind when designing randomized evaluations with a view to identifying policies that will work at scale? Perhaps the key point is to remember what small pilot experiments are good for and what they are not good for. Formulation of a successful large-scale public policy begins with the identification of a promising concept, which requires elaborating a model of the mechanism at play. Small-scale experiments can identify these concepts, both by pinpointing the sources of specific problems and testing approaches of dealing with them. Fully understanding the key mechanism behind successful interventions is often likely to take more than one experiment. In the case of education, early experiments by Glewwe, Kremer, and Moulin (2009) and the initial *balsakhi* results (Banerjee, Cole, Duflo, and Linden 2007) helped identify the core problem of the mismatch between what gets taught and what the children need to learn, but the results could have been explained by other factors (for example, in the *balsakhi* study, class size went down, and the instructor was younger and closer to the students). Based on this work, Duflo, Dupas, and Kremer (2011) designed an experiment that specifically investigated the potential of matching children by level, disentangling it from the effect of being assigned different kinds of teachers (for example, those who may be closer to the students and have better incentives), and found that it indeed matters. If the objective is to design or test a model, the researcher can ignore most of the concerns that we talked about in this paper. Something valuable will be learnt anyway. This is the equivalent of what is sometimes called "stage one" in venture capital investing.[6]

It would of course be dangerous to advocate a policy scale-up based exclusively on results from an investigation of this sort. The importance of all the issues we discussed earlier needs to be evaluated, which typically requires additional experimental work. We now consider them in turn (though not in the order in which we first discussed them).

Context dependence can be assessed by replications, either of the same experiments or of related experiments (that is, by experiments that test programs inspired by the same general idea). To assess whether a program is ready to be scaled up, or should be evaluated again first (perhaps starting on a smaller scale), policymakers

[6] This staged approach, inspired by venture capital funding process, is now explicitly adopted by some aid organizations, such as the US Agency for International Development's Development Ventures and the Global Innovation Fund (US Agency for International Development 2017; Global Innovation Fund 2017).

should ideally be able to rely on an aggregation of all the existing reliable evidence, randomized or not. Many are skeptical as to whether needed replications would be undertaken, but this skepticism does not seem warranted. With the proliferation of experiments in the last decade or so, there starts to be a critical mass of work on many key issues. Programs that appear to be particularly promising are more likely to be replicated. For example, Banerjee, Duflo et al. (2015) present six separate evaluations of an asset transfer program developed by the Bangladeshi Rural Advancement Committee that is being implemented around the world.

Once a program has passed the proof-of-concept test and is chosen to be scaled up, the next step is to develop an implementable large-scale policy version, and to subject it to a stage-two trial, meaning a larger trial that will confront and document the problems that the program would have at scale.[7] Designing this intervention typically requires combining an understanding of the mechanism underlying the concept with insight into how the particular government (or other large-scale implementer) works as an organization, which we have referred to elsewhere as getting "inside the machine" (Banerjee 2007) or as fixing the "plumbing" (Duflo 2017). For such trials to be informative, a number of critical design issues need to be addressed, which is what we turn to next.

To address the possibility of *randomization bias*, the organization that implements a stage-two trial must be the organization that will eventually implement it at scale, if it were to be scaled up. Within this organization, it must be implemented by the regular staff, not by a group of external experts. It also needs to be run in an area that is representative of where it would be scaled up eventually. For example, Muralidharan and Sundararaman (2015) randomly chose districts to run their market-level private voucher experiments.

For researchers, a strong temptation in a stage-two trial will be to do what it takes "to make it work," but the risk of *implementation challenges* means that it is important to think about how far to go in that direction. On the one hand, trial and error will be needed to embed any new intervention within an existing bureaucracy. Anything new is challenging, and at the beginning of a stage-two trial, considerable time needs to be spent to give the program a fair shot. On the other hand, if the research team embeds too much of its own staff and effort and ends up substituting for the organization, not enough will need to be learnt about where implementation problems might emerge. Our suggestion is to pilot a potential implementation system with some external staff support initially, and then to move progressively towards a more hands-off approach, but to continue to monitor processes carefully in at least a representative sample of locations.

When an intervention that can work at scale in the right organization has been successfully developed, it can be deployed at scale to evaluate the full effect of the intervention, including any spillover and market-level effects. A number of studies have been designed to estimate *equilibrium effects* by randomizing at the level of the

[7] In some cases, it will make sense to go straight to a fairly large stage-two trial, because the experiment does not even make sense on a small scale.

relevant market. Theory (and common sense) can guide the key design questions: On what variables (if any) do we expect to see equilibrium effects? What is the nature of those effects? Are we moving down a demand curve? Do they arise because of competition? What is the relevant market?

There are situations where relevant market equilibrium effects cannot be experimentally estimated—for example, because the entire country would be the right market. In the case of free secondary schooling in Ghana, for example, we expect that secondary school graduates will have a national market essentially (they can move to Accra, and they compete nationally for teacher and nurse training slots). In that case, the best a researcher can do is to combine the partial equilibrium results with some modeling and known elasticities, and exploit the understanding of the context to make some predictions about possible market equilibrium. In Ghana, the cohort that was subject to the experiment of Duflo, Dupas, and Kremer (2017) graduated as part of a "double cohort" (because the length of secondary school was brought down from four to three years after this cohort matriculated). Therefore, the authors conclude that the partial equilibrium impacts within that double cohort are probably a good approximation of what would happen if free secondary school doubled the share of graduates, at least in the short run. It is also useful to try to identify situations where one would expect market equilibrium effects to be too small to matter (consider, for example, the case of a preschool math programs that can be run by existing teachers).

Although conceptually distinct, *spillover effects* can be evaluated experimentally in the same way as market equilibrium, by randomizing at the appropriate level (and randomizing in two steps if one is particularly interested in the spillover themselves, not just the total effect). One open issue that has proven difficult is the identification of nonlinear spillovers. Conceptually, this requires randomization of treatment intensity at several points and then comparison of treated and untreated units in each treatment intensity. Crepon et al. (2013) adopts this design in their experiment on the French labor market (they treat 25, 50, and 75 percent of the units). Similarly, Banerjee et al. (2014) treat 25, 50, 75, or 100 percent of the police officers in police stations in Rajasthan. In practice, the Crepon et al. (2013) study lacks enough statistical precision to identify differential spillover effects (despite working at the scale of half of France). Banerjee et al. (2014) find a nonlinearity in overall effect of the treatment (there is no impact in treating 25 percent of the police officers, and the effect is the same when 50, 75, and 100 percent of the officers are trained), but they do not specifically track spillovers.

Finally, implementing the scale-up with the organization that would finally implement, within their standard operating procedures, and at a scale sufficient to detect market equilibrium effects will also give a chance for any potential *political backlash* to manifest itself. As mentioned above, this happened in the Banerjee, Duflo, Imbert, Mathew, and Pande (2017) study of anticorruption reforms in India. When backlash happens, it is worth exploring whether some changes in potentially inessential program details (perhaps some side payments to the aggrieved parties) are available. It is also important to try to anticipate the backlash and create a

constituency for the reform from the start. Finally, the potential for political backlash may provide an argument for not doing too many pilots, since large-scale programs are less likely to be scotched.

■ *Thanks to Richard McDowell, Harris Eppsteiner, and Madeline Duhon for research assistance; to Tamayata Bansal, Sugat Bajracharya, Anupama Deshpande, Blaise Gonda, John Firth, Christian Larroulet, Adrien Lorenceau, Jonathan Mazumdar, Manaswini Rao, Paribhasha Sharma, Joseph Shields, Zakaria Siddiqui, Yashas Vaidya, and Melanie Wasserman for field management; to Diva Dhar for supervision; and to Shaher Bhanu Vagh for the educational test design and analysis. Special thanks to the staff of Pratham for their openness and engagement, and to the William and Flora Hewlett Foundation, the International Initiative for Impact Evaluation, the Government of Haryana, and the Regional Centers for Learning on Evaluation and Results for their financial support and commitment.*

References

Akram, Agha Ali, Shyamal Chowdhury, and Ahmed Musfiq Mobarak. 2017. "Effects of Emigration on Rural Labor Markets." Unpublished paper

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130(3): 1117–65.

Andrabi, Tahir, Jishnu Das, Selcuk Ozyurt, and Niharika Singh. 2017. "Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools." Unpublished paper.

Angelucci, Manuela, and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *American Economic Review* 99(1): 468–508.

Angrist, Joshua, and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Athey, Susan, and Guido Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, 323–85. ScienceDirect.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2011. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA." *Review of Economic Studies* 79(1): 37–66.

Banerjee, Abhijit. 2007. *Making Aid Work.* Cambridge, MA: MIT Press.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." NBER Working Paper 22746.

Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in India." *American Economic Journal: Economic Policy* 2(1): 1–30.

Banerjee, Abhijit, Sharon Barnhardt, and Esther Duflo. 2015. "Movies, Margins and Marketing: Encouraging the Adoption of Iron-Fortified Salt." NBER Working Paper 21616.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2017. "Decision Theoretic Approaches to Experiment Design and External Validity." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo,

141–74. ScienceDirect.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2014. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training." NBER Working Paper 17912.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3): 1235–64.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.*

Banerjee, Abhijit, Esther Duflo, and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Public Health Care System." *Journal of the European Economic Association* 6(2–3): 487–500.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Pariente, 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).

Banerjee, Abhijit, Esther Duflo, Clement Imbert, Santhosh Mathew, and Rohini Pande. 2017. "E-governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India." CEPR Discussion Paper 1176.

Banerjee, Abhijit, Esther Duflo, and Michael Kremer. Forthcoming. "The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy." In *The State of Economics, The State of the World*, edited by Kaushik Basu.

Banerjee, Abhijit, Rema Hanna, Jordan Kyle, Benjamin A. Olken, and Sudarno Sumarto. 2016. "Tangible Information and Citizen Empowerment: Identification Cards and Food Subsidy Programs in Indonesia." https://economics.mit.edu/files/11877.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7(1): 1–21.

Banerji, Rukmini, Madhav Chavan, and Usha Rane. 2004. "Learning to Read." *India Seminar*, April. http://www.indiaseminar.com/2004/536/536%20rukmini%20banerji%20%26%20et%20al.htm.

Barrera-Osorio, Felipe, and Leigh L. Linden. 2009. "The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia." World Bank Policy Research Working Paper 4836.

Berry, James, Greg Fischer, and Raymond P. Guiteras. 2015. "Eliciting and Utilizing Willingness-to-Pay: Evidence from Field Trials in Northern Ghana." CEPR Discussion Paper 10703.

Blair, Graeme, Radha K. Iyengar, and Jacob N. Shapiro. 2013. "Where Policy Experiments Are Conducted in Economics and Political Science: The Missing Autocracies." http://scholar.princeton.edu/sites/default/files/jns/files/blair_iyengar_shapiro_rcts_20may13.pdf.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2015. "Interventions and Institutions: Experimental Evidence on Scaling up Education Reforms in Kenya." Unpublished paper.

Breza, Emily, and Cynthia Kinnan. 2016. "Measuring the Equilibrium Impacts of Credit: Evidence from the Indian Microfinance Crisis." http://faculty.wcas.northwestern.edu/~cgk281/Eqm_impacts_credit.pdf.

Buera, Francisco, Joseph Kaboski, and Yongseok Shin. 2012. "The Macroeconomics of Microfinance." NBER Working Paper 17905.

Card, David, Jochen Kluve, and Andrea Weber. 2010. "Active Labour Market Policy Evaluations: A Meta-Analysis." *Economic Journal* 120(548): F452–77.

Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg. 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *American Economic Review* 102(4): 1279–1309.

Clasen, Thomas F., Kelly T. Alexander, David Sinclair, Sophie Boisson, Rachel Peletz, Howard H. Chang, Fiona Majorin, and Sandy Cairncross. 2015. "Interventions to Improve Water Quality for Preventing Diarrhoea." *Cochrane Database of Systematic Reviews* 10.

Coalition for Evidence-Based Policy. 2013. *Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive versus Weak or No Effects.* Washington, DC: Coalition for Evidence-Based Policy.

Cohen, Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125(1): 1–45.

Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128(2): 531–80.

Cunha, Jesse, Giacomo De Giorgi, and Seema Jayachandran. 2011. "The Price Effects of Cash

versus In-Kind Transfers." NBER Working Paper 17456.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424–55.

Duflo, Esther. 2017. "The Economist as Plumber." NBER Working Paper 23213.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5): 1739–74.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives and Pupil–Teacher Ratios." *Journal of Public Economics* 123: 92–110.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2017. "The Impact of Free Secondary Education: Experimental Evidence from Ghana." https://web.stanford.edu/~pdupas/DDK_GhanaScholarships.pdf.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School. " *American Economic Review* 102(4): 1241–78.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118(3): 815–42.

Dupas, Pascaline. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82(1): 197–228.

Dupas, Pascaline, and Edward Miguel. 2017. "Impacts and Determinants of Health Levels in Low-Income Countries." In *Handbook of Economic Field Experiments*, edited by Abhijit V. Banerjee and Esther Duflo, 3–93. ScienceDirect.

Evidence Action. 2014. "Kenya Deworming Results Announced: 6.4 Million Children Worm-Free and Ready to Learn." *Evidence Action*, August 4. http://www.evidenceaction.org/blog-full/kenya-deworming-results-announced-6-4-million-children-worm-free-and-ready-to-learn.

Finn, Jeremy D., and Charles M. Achilles. 1990. "Answers and Questions about Class Size: A Statewide Experiment." *American Educational Research Journal* 27(3): 557–77.

Glennerster, Rachel. 2017. "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, 175–243. ScienceDirect.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1(1): 112–35.

Global Innovation Fund. 2017. *Stages of Financing.* Washington, DC: Global Innovation Fund.

Hausman, Jerry A., and David A. Wise. 1985. *Social Experimentation.* University of Chicago Press.

Heckman, James. 1992. "Randomization and Social Programs." In *Evaluating Welfare and Training Programs*, edited by Charles Manski and Irwin Garfinkel. Cambridge, MA: Harvard University Press.

Heckman, James J., Lance Lochner, and Christopher Taber. 1998. "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics* 1: 1–58.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Innovations for Poverty Action. 2015. *Targeted Lessons to Improve Basic Skills.* New Haven, CT: Innovations for Poverty Action.

International Initiative for Impact Evaluation. 2017. "Systematic Reviews." International Initiative for Impact Evaluation. http://www.3ieimpact.org/en/evidence/systematic-reviews/ (accessed June 27, 2017).

Kapur, Avani, and Lorenza Icaza. 2010. "An Institutional Study of Read India in Bihar and Uttarakhand." Unpublished paper.

Kremer, Michael, and Rachel Glennerster. 2011. "Improving Health in Developing Countries: Evidence from Randomized Evaluations." In *Handbook of Health Economics*, edited by Mark V. Pauly, Thomas G. McGuire, and Pedro P. Barros, 201–315. ScienceDirect.

Lengeler, Christian. 2004. "Insecticide-Treated Bed Nets and Curtains for Preventing Malaria." *Cochrane Database of Systematic Reviews* 2.

Manski, Charles F., and Irwin Garfinkel. 1992. *Evaluating Welfare and Training Programs.* Cambridge, MA: Harvard University Press.

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments." *Review of Educational Research* 85(3): 353–94.

McKenzie, David, and Susana Puerto. 2017. "Growing Markets through Business Training for Female Entrepreneurs." World Bank Policy Research Working Paper 7793.

Meager, Rachael. 2016. "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomized Experiments." https://economics.mit.edu/files/11443.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159–217.

Muralidharan, Karthik. 2017. "Field Experiments in Education in Developing Countries." In *Handbook of Economic Field Experiments*, edited by Abhijit V. Banerjee and Esther Duflo, 323–85. ScienceDirect.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106(10): 2895–2929.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics* 130(3): 1011–66.

Newhouse, Joseph P. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, MA: Harvard University Press.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments when Context Matters." *American Economic Review* 105(5): 471–75.

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Education and Behavioral Statistics* 6(4): 377–401.

Sackey, Ken. 2017. "Free SHS to Commence September 2017—President Akufo-Addo." *Ghana News Agency*, February 11. http://www.ghananewsagency.org/education/free-shs to-commence-september-2017-president-akufo-addo-113179.

Sandefur, Justin. 2011. "Held Hostage: Funding for a Proven Success in Global Development on Hold in Kenya." *Center for Global Development*, April 25.

Schweinhart, Lawrence J., Helen V. Barnes, and David P. Weikart. 1993. *Significant Benefits: The High/Scope Perry Preschool Study through Age 27.* Ypsilanti, MI: High/Scope Press.

Sharma, Paribhasha, and Anupama Deshpande. 2010. "Teachers' Perception of Primary Education and Mothers' Aspirations for Their Children—A Qualitative Study in Bihar and Uttarakhand." Unpublished paper.

Tarozzi, Alessandro, Aprajit Mahajan, Brian Blackburn, Dan Kopf, Lakshmi Krishnan, and Joanne Yoong. 2014. "Micro-loans, Insecticide-Treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India." *American Economic Review* 104(7): 1909–41.

Taylor-Robinson, David C., Nicola Maayan, Karla Soares-Weiser, Sarah Donegan, and Paul Garner. 2015. "Deworming Drugs for Soil-Transmitted Intestinal Worms in Children: Effects on Nutritional Indicators, Haemoglobin, and School Performance." *Cochrane Database of Systematic Reviews* 7.

Townsend, Robert. 2010. "Financial Structure and Economic Welfare: Applied General Equilibrium Development Economics." *Annual Review of Economics* 2(1): 507–46.

US Agency for International Development (USAID). 2017. "Development Innovation Ventures." Webpage. https://www.usaid.gov/div (accessed June 27, 2017).

Vivalt, Eva. 2016. "How Much Can We Generalize from Impact Evaluations?" http://evavivalt.com/wp-content/uploads/2014/12/Vivalt_JMP_latest.pdf.

This article has been cited by:

1. Alex Eble, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Tony Brady, Piotr Gawron, Stijn Vansteelandt, Peter Boone, Diana Elbourne. 2021. How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia. *Journal of Development Economics* **148**, 102539. [Crossref]
2. Burt S. Barnow, David H. Greenberg. 2020. Conducting Evaluations Using Multiple Trials. *American Journal of Evaluation* **41**:4, 564-580. [Crossref]
3. Nur Cahyadi, Rema Hanna, Benjamin A. Olken, Rizal Adi Prima, Elan Satriawan, Ekki Syamsulhakim. 2020. Cumulative Impacts of Conditional Cash Transfer Programs: Experimental Evidence from Indonesia. *American Economic Journal: Economic Policy* **12**:4, 88-110. [Abstract] [View PDF article] [PDF with links]
4. Omar Al-Ubaydli, John A. List, Dana Suskind. 2020. 2017 KLEIN LECTURE: THE SCIENCE OF USING SCIENCE: TOWARD AN UNDERSTANDING OF THE THREATS TO SCALABILITY. *International Economic Review* **61**:4, 1387-1409. [Crossref]
5. Diane Alexander. 2020. How Do Doctors Respond to Incentives? Unintended Consequences of Paying Doctors to Reduce Costs. *Journal of Political Economy* **128**:11, 4046-4096. [Crossref]
6. Tamara McGavock. 2020. Here waits the bride? The effect of Ethiopia's child marriage law. *Journal of Development Economics* 102580. [Crossref]
7. Tahir Andrabi, Jishnu Das, Asim I. Khwaja, Selcuk Ozyurt, Niharika Singh. 2020. Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools. *American Economic Review* **110**:10, 3315-3349. [Abstract] [View PDF article] [PDF with links]
8. Abhijit Banerjee, Esther Duflo, Clément Imbert, Santhosh Mathew, Rohini Pande. 2020. E-governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India. *American Economic Journal: Applied Economics* **12**:4, 39-72. [Abstract] [View PDF article] [PDF with links]
9. Elia Axinia Machado, Helene Purcell, Andrew M. Simons, Stephanie Swinehart. 2020. The Quest for Greener Pastures: Evaluating the Livelihoods Impacts of Providing Vegetation Condition Maps to Pastoralists in Eastern Africa. *Ecological Economics* **175**, 106708. [Crossref]
10. Di Mo, Yu Bai, Yaojiang Shi, Cody Abbey, Linxiu Zhang, Scott Rozelle, Prashant Loyalka. 2020. Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China. *Journal of Development Economics* **146**, 102487. [Crossref]
11. Catalina Herrera-Almanza, Maria F. Rosales-Rueda. 2020. Reducing the Cost of Remoteness: Community-Based Health Interventions and Fertility Choices. *Journal of Health Economics* **73**, 102365. [Crossref]
12. Cristina Corduneanu-Huci, Michael T. Dorsch, Paul Maarek. 2020. The politics of experimentation: Political competition and randomized controlled trials. *Journal of Comparative Economics* . [Crossref]
13. William F. Shughart, Diana W. Thomas, Michael D. Thomas. 2020. Institutional Change and the Importance of Understanding Shared Mental Models. *Kyklos* **73**:3, 371-391. [Crossref]
14. Brahm Fleisch. 2020. Twenty five years of the Journal of Educational Change: A Perspective from the Global South. *Journal of Educational Change* **21**:3, 479-483. [Crossref]
15. OMAR AL-UBAYDLI, MIN SOK LEE, JOHN A. LIST, CLAIRE L. MACKEVICIUS, DANA SUSKIND. 2020. How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy* **3**, 1-48. [Crossref]

16. GLENN W. HARRISON. 2020. Field experiments and public policy: festina lente. *Behavioural Public Policy* **51**, 1-8. [Crossref]
17. TECK-HUA HO, CHING LEONG, CATHERINE YEUNG. 2020. Success at scale: six suggestions from implementation and policy sciences. *Behavioural Public Policy* 1-9. [Crossref]
18. Esther Duflo. 2020. Field Experiments and the Practice of Policy. *American Economic Review* **110**:7, 1952-1973. [Citation] [View PDF article] [PDF with links]
19. Emma Gibbs, Charlotte Jones, Jess Atkinson, Ian Attfield, Rona Bronwin, Rachel Hinton, Amy Potter, Laura Savage. 2020. Scaling and 'systems thinking' in education: reflections from UK aid professionals. *Compare: A Journal of Comparative and International Education* **20**, 1-20. [Crossref]
20. Qingyang Huang, Chang Liu, Li-An Zhou. 2020. Farewell to the God of Plague: Estimating the effects of China's Universal Salt Iodization on educational outcomes. *Journal of Comparative Economics* **48**:1, 20-36. [Crossref]
21. James Berry, Harini Kannan, Shobhini Mukherji, Marc Shotland. 2020. Failure of frequent assessment: An evaluation of India's continuous and comprehensive evaluation program. *Journal of Development Economics* **143**, 102406. [Crossref]
22. Ambrish Dongre, Vibhu Tewary. 2020. Pain without gain?: Impact of school rationalisation in India. *International Journal of Educational Development* **72**, 102142. [Crossref]
23. Noam Angrist, Peter Bergman, Caton Brewster, Moitshepi Matsheng. 2020. Stemming Learning Loss During the Pandemic: A Rapid Randomized Trial of a Low-Tech Intervention in Botswana. *SSRN Electronic Journal* . [Crossref]
24. Michael Dorsch, Paul Maarek. 2020. The Politics of Experimentation: Political Competition and Randomized Controlled Trials. *SSRN Electronic Journal* . [Crossref]
25. Juan E. Saavedra, Emma Näslund-Hadley, Mariana Alfonso. 2019. Remedial Inquiry-Based Science Education: Experimental Evidence From Peru. *Educational Evaluation and Policy Analysis* **41**:4, 483-509. [Crossref]
26. Joel Slemrod. 2019. Tax Compliance and Enforcement. *Journal of Economic Literature* **57**:4, 904-954. [Abstract] [View PDF article] [PDF with links]
27. Chris Cunningham, Solveig A. Cunningham, Nafisa Halim, Kathryn M. Yount. 2019. Public Investments in Education and Children's Academic Achievements. *The Journal of Development Studies* **55**:11, 2365-2381. [Crossref]
28. Eszter Czibor, David Jimenez-Gomez, John A. List. 2019. The Dozen Things Experimental Economists Should Do (More of). *Southern Economic Journal* **86**:2, 371-432. [Crossref]
29. Timothy G. Evans. 2019. Driving Demand and Delivery for Mental Health. *American Journal of Public Health* **109**:S3, S164-S164. [Crossref]
30. Jake Bowers, Paul F. Testa. 2019. Better Government, Better Science: The Promise of and Challenges Facing the Evidence-Informed Policy Movement. *Annual Review of Political Science* **22**:1, 521-542. [Crossref]
31. Lisa Cameron, Susan Olivia, Manisha Shah. 2019. Scaling up sanitation: Evidence from an RCT in Indonesia. *Journal of Development Economics* **138**, 1-16. [Crossref]
32. Sam Harper. 2019. Comment on the equity impact of women's community groups on inequalities in neonatal mortality. *International Journal of Epidemiology* **48**:1, 182-185. [Crossref]
33. Eszter Czibor, David Jimenez-Gomez, John A. List. 2019. The Dozen Things Experimental Economists Should Do (More Of). *SSRN Electronic Journal* . [Crossref]
34. Thomas Le Barbanchon, Diego Ubfal, Federico Araya. 2019. The Effects of Working While in School: Evidence from Uruguayan Lotteries. *SSRN Electronic Journal* . [Crossref]

35. Yanying Chen, Yi Jin Tan. 2018. The effect of non-contributory pensions on labour supply and private income transfers: evidence from Singapore. *IZA Journal of Labor Policy* **7**:1. . [Crossref]
36. Moshe Justman. 2018. Randomized controlled trials informing public policy: Lessons from project STAR and class size reduction. *European Journal of Political Economy* **54**, 167-174. [Crossref]
37. Jörg Peters, Jörg Langbein, Gareth Roberts. 2018. Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer* **33**:1, 34-64. [Crossref]
38. Qingyang Huang, Chang Liu, Li-An Zhou. 2018. Farewell to the God of Plague: Estimating the Effects of Universal Salt Iodization on School Enrollment. *SSRN Electronic Journal* . [Crossref]
39. Stephanie Moulton, J. Michael Collins, Olga Kondratjeva. 2018. Pragmatic Field Experiments in Policy Research: The Case of a Pilot Program for Municipal Water Customers. *SSRN Electronic Journal* . [Crossref]
40. Faraz Usmani, Marc Jeuland, Subhrendu K. Pattanayak. 2018. NGOs and the Effectiveness of Interventions. *SSRN Electronic Journal* . [Crossref]
41. Tahir Andrabi, Jishnu Das, Asim Ijaz Khwaja, Selcuk Ozyurt, Niharika Singh. 2018. Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools. *SSRN Electronic Journal* . [Crossref]
42. D. Brent Edwards. Critically Understanding Impact Evaluations: Technical, Methodological, Organizational, and Political Issues 23-67. [Crossref]
43. Mariella Gonzales, Gianmarco León, Luis Martinez. 2018. Monetary Incentives to Vote: Evidence From a Nationwide Policy. *SSRN Electronic Journal* . [Crossref]
44. Rob Bauer, Inka Eberhardt, Paul Smeets. 2017. Financial Incentives Beat Social Norms: A Field Experiment on Retirement Information Search. *SSRN Electronic Journal* . [Crossref]