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[1]

September 2016

Abstract

The promise of randomized controlled trials (RCTs) is that evidence gathered through the evaluation of a specific program helps us—possibly after several rounds of fine-tuning and multiple replications in different contexts—to inform policy. However, critics have pointed out that a potential constraint in this agenda is that results from small, NGO-run "proof-of-concept" studies may not apply to policies that can be implemented by governments on a large scale. After discussing the potential issues, this paper describes the journey from the original concept to the design and evaluation of scalable policy. We do so by evaluating a series of strategies that aim to integrate the NGO Pratham's "Teaching at the Right Level" methodology into elementary schools in India. The methodology consists of re-organizing instruction based on children's actual learning levels, rather than on a prescribed syllabus, and has previously been shown to be very effective when properly implemented. We present RCT evidence on the designs that failed to produce impacts within the regular schooling system but helped shape subsequent versions of the program. As a result of this process, two versions of the programs were developed that successfully raised children's learning levels using scalable models in government schools.

[1] Banerjee: MIT, NBER, and BREAD, banerjee@mit.edu; Banerji: ASER Centre and Pratham, rukmini.banerji@pratham.org; Berry: Cornell University, jimberry@cornell.edu, Duflo: MIT, NBER, and BREAD, eduflo@mit.edu, Kannan: J-PAL, harini.kannan@ifmr.ac.in, Mukherji: J-PAL, shobhini.mukerji@ifmr.ac.in, Shotland: J-PAL, shotland@mit.edu, Walton: Harvard University, michael_walton@hks.harvard.edu. Thanks to Richard Mc- Dowell, 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 Foun- dation, 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.

1 Introduction

Randomized controlled trials (RCTs) have been used in economics and other social sciences for decades,[2] and 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 RCTs are particularly credible from the point of view of internal validity (Athey and Imbens, 2016). 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.[3] In particular, they argue that it is not at all clear that results from small, NGO-run "proof-of-concept" studies should be directly turned into recommendations for policies for implementation by governments on a large scale (Deaton 2010).[4]

In this paper we explore the challenges of going from a single localized RCT to a policy implemented at scale, illustrated with the example of an educational intervention that in fact successfully traversed the distance from an NGO implemented pilot in a few slums to a policy implemented at scale by state governments in India.[5] We demonstrate that external validity is neither taken for granted nor unattainable. Rather, the journey from internal to external validity is a process that involves trying to identify the underlying mechanisms, possibly refining the

[2] Early examples of such studies include the Negative Income Tax experiments (Hausman and Wise 1985), the RAND Health Insurance Experiment (Newhouse 1993), a series of welfare reform experiments in the 1980s and 1990s (Manski and Garfinkel 1992), work on education (such as the Perry Pre-School Project and Project STAR), and studies in criminology (such as the Minneapolis Domestic Violence Experiment, Berk and Sherman 1984).
[3] Some critics go further, suggesting there is an explicit tradeoff: whatever gains RCTs make in internal validity, they lose in external validity when applied to different contexts. This argument, as should be obvious, relies on the implicit assumption that the identifying variation in large studies covers many locations, which is not necessarily true.
[4] That said, many pilots these days are enormous, covering many millions of people. We will discuss one such pilot briefly below.
[5] States in India are the size of countries in Europe.

intervention model based on the understanding of the mechanisms and other practical considerations, and often requiring multiple iterations of experimentation.

2 From proof of concept to scalable policies: The challenges

Just as for efficacy studies in medical trials, which are usually performed in tightly controlled conditions inside of the lab, it often makes sense to verify the proof of concept of a new social program under ideal conditions—through finding a context and implementation partner most likely to make the model work.[6] However, when a potential program is tested on a small scale, the results, while informative, need not be good predictors of what would happen if a similar policy were to be implemented on a large scale. It is not uncommon for studies to fail to replicate results that had been established in smaller RCTs elsewhere.

There are various obstacles:

Equilibrium Effects. When an intervention is implemented at scale, it could change the nature of the market, and consequently, the effect of the scaled-up program may be nothing like the effect of the program in its small-scale implementation. For example, Heckman, Lochner and Taber (1998) argue, based on model simulations, that education interventions that produce large increases in educational attainment may thereby decrease the overall return to education. Of course one could also imagine situations where ignoring the equilibrium effect leads to an underestimation of the treatment effect. For example, ignoring the possibility that an intervention targeted to some children in a school also benefits others in the school who were in the control group will give us a treatment effect that is too small. This might occur through adjustments in teaching within the school or peer effects, for example.

To take account of the possibility of equilibrium effects several recent papers employ a two-

[6] Classing, Pedro-i-Miguel and Snowberg (2012) provide a formal justification of this argument.

stage randomization procedure in which the treatment is randomly assigned at the market level in addition to the random assignment within a market.[7] Using such a design, Crepon et al. (2013) find evidence of equilibrium effects in a job placement assistance program in France. The experiment varied the treatment density 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.

On the other hand, Muralidharan and Sundararaman (2015) who adopt a similar design to evaluate a school voucher program in Andhra Pradesh, India, find no evidence of equilibrium effects. Villages were first assigned to treatment and control groups, and within the treatment group, individuals were randomly assigned to receive a voucher. The authors find higher test scores for individuals who received a voucher in some subjects (but no improvement in core competencies) but the comparison of individuals who did not receive vouchers in the treatment villages with individuals in control villages shows no evidence of market-level externalities of the voucher program. It is worth noting that the equilibrium effect in this case could have had effects in either direction—those who did not receive vouchers in treatment villages may be worse off than in control because the private schools were more crowded or better off than in control because the public schools were less crowded.

Miguel and Kremer (2004) randomized the assignment of medicines for deworming just at the school level. However, they then 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 medicine on those who did not themselves take it. They find large

[7] This procedure has been implemented previously to examine peer effects in program take up (Duflo and Saez 2003).

positive spillover effects, which tell us that the naïve comparison of treatment and control schools is biased downwards.

A number of other recent experiments were designed to estimate only the full equilibrium effect, by conducting the randomization only at the market level. Muralidharan et al. (2015) evaluate the rollout of an electronic payments system for the NGREGS 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 authors document that 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 2010); Angelucci and De Giorgi 2009; Attanasio, Meghir, and Santiago 2011).

Political Reactions. A variant of an equilibrium effect is the potential political reaction to scaled programs. For example, political resistance to or support for a program may build up when the program reaches a sufficient scale. Alternatively, corrupt officials may be more likely to exploit 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 RCTs, 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 caused donors to cut off education aid, including support for the deworming program. The program ultimately re-started in 2012 (Sandefur, 2011; Evidence Action, 2016).

Banerjee, Duflo, et al. (2016) provide an interesting example of political backlash leading to the demise of a promising anti-corruption program in India. The program was randomized at the Gram Panchayat level, but even though it was meant to be a pilot, it covered (half of) 12

districts, almost 3000 GPs and 33 million people. The GP officials and their immediate superiors at the block or district level (there are 16 blocks on average in a district, and 15 GPs per block) were dead set against the intervention for the simple reason that it threatened their rents and opposed the implementation of the program (the experiment did find a significant decline in rent-seeking and the wealth of program officials). Ultimately, in part due to the fact that the reduction in corruption could not be proven until a household survey was completed, these officials were successful in lobbying the State government, and the intervention was cancelled.[8]

This is an example of a pilot that was much larger than the typical proof of concept study, so much so that the group it took on was large enough to have political influence. A smaller pilot might have had a less difficult time since its opponents would have been less numerous and therefore less powerful, but this effect would have been missed. At the same time, being a pilot and therefore subject to review did not help—it made it vulnerable to being shut down, which is what happened.

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

Replication of experiments allows us to say something about context dependence. Systematic reviews (such as Cochrane Reviews) bring evidence from replications together in one place, and their use is expanding in economics. The International Initiative for Impact Evaluation maintains a database of systematic reviews of impact evaluations in developing countries that contains 303 studies as of this writing. Cochrane reviews have been compiled on topics such as water quality

[8] Although the evaluation did help: based on the results (which came out after the program was cancelled in Bihar), the intervention was extended to the same workfare program in most other States, and there are discussions to extend it to other government transfers programs.

interventions (Clasen et al. 2015), mosquito nets (Lengeler, 2004), and deworming of schoolchildren (Taylor-Robinson et al. 2015). While these reviews reveal a surprising amount of homogeneity across studies, we need to know much more. 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 will hopefully allow similar analyses across many more programs.

Several recent studies and journal volumes compile the results from multiple interventions in the same publication. A special issue of the *American Economic Journal: Applied Economics* is devoted to 6 experimental studies of microfinance. Although these studies were not conducted in coordination with one another, the overall conclusions are quite consistent across studies: the interventions showed modest increases in business activity but limited evidence for increases in consumption (Banerjee, Karlan, and Zinman, 2015).

However as argued by Banerjee, Chassang, and Snowberg (2016), prior-free extrapolation is not possible. To aggregate these effects, one has to start from some assumption about the potential distribution of treatment effects. In the economics literature, this is often done informally (Vivalt, 2015), 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 used retrospective data, and 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

assume 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.[9]

It is worth noting however that 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 both the successes and the failures (rather than just letting the failures cancel out the successes). The theory can then have other predictions that could be tested in future experiments, and all of that could feed into the design of the scaled up intervention.

Banerjee and Duflo provide some informal examples of how this may be done in their book Poor Economics (2011). Kremer and Glennerster (2012) develop an interpretative framework for the very high price sensitivity results found in RCTs on the take up of preventive healthcare. They propose a number of number alternative theories featuring liquidity constraints, lack of information, non-monetary costs, or behavioral biases such as present bias and limited attention. Dupas and Miguel (2016) provide an excellent summary of what the evidence from RCTs (mostly subsequent to the Kremer and Glennerster paper) tell us about the plausibility of each of these theories and argue that the subsequent evidence supports some aspects of the Kremer-Glennerster framework and rejects others. The point is that many of those experiments were designed precisely with that framework in mind, which makes them much more informative. We will return to the role of developing a theory in the scale up process in the next section.

[9] McEwan (2015) is another example of meta-analysis. He analyzes the results of 77 RCTs 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.

Randomization bias (or site-selection bias) (Heckman, 1992). Organizations or individuals who agree to participate in an early experiment may be different from the rest of the population; they may be more suited to the intervention or particularly motivated to see the program succeed.

Glennerster (2016) lists the characteristics of a good partner to work with for an RCT, and it is evident that many organizations in developing countries do not meet their criteria. For example, senior staff must be open to the possibility of the program not working and be willing to have these results publicized.[10] They also must be prepared to spend time organizing the randomized implementation and ensure that program implantation follows the randomized design, providing relatively uniform implementation in the treatment group while not contaminating control group. These characteristics may be related to strong program implementation and lead to larger effect sizes than those in the scaled program, if it is run by a less stellar organization.

Site-selection bias can manifest itself when NGOs with an interest in detecting impact choose locations with the most need, and the population receiving the intervention in the RCT are likely to have larger impacts than those in a scaled intervention. In other words, NGOs can leverage context dependence to maximized detected impacts. For example, Banerjee, Duflo, and Barnhardt (2015) find no impact on anemia of free iron-fortified salt, in contrast with previous RCTs which led to the approval of the product for general marketing. Part of the reason is that there was no specific effort to make the treated population use the salt. But even the treatment on treated estimates are lower in their study, except for one group: adolescent women. And young women were precisely the groups that were targeted in previous studies.

Another related issue is that in interventions in which individuals select into treatment, RCTs

[10] Brigham et al. (2013) conduct an experiment soliciting microfinance institutions for randomized evaluations, providing priming for either positive results or null results. The authors find that priming for positive results has significant impacts on interest in an RCT.

may induce different groups of individuals to select into the experiment (and undertake the treatment) than those affected by the scaled intervention. If treatment effects are heterogeneous across these groups, the estimated effect from the RCT may not apply to a broader population (see, e.g., Heckman and Vytlacil 2007).[11]

Several recent papers examine issues of randomization bias across a large number of RCTs. Vivalt (2015) compiles data from over 400 RCTs and examines the relationship between effect size and study characteristics. Studies evaluating NGO or researcher-run interventions tend to have higher point estimates than RCTs run with governments, as do studies with smaller sample sizes. Allcott (2015) presents the results of 111 RCTs of the Opower program in which households are presented with information on energy conservation and energy consumption of neighbors. He finds that the first 10 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 RCTs across countries and find that RCTs are disproportionally conducted in countries with democratic governments.[12]

Piloting bias/implementation challenges at scale. For an intervention into a government program to go to scale, large parts of an entire bureaucracy have to buy into and adopt it. The intense monitoring that is possible in a pilot may no longer be feasible when that happens, and even when it's possible may require a special effort. For example, schooling reforms often require full buy-in from teachers and school principals in order to be effective: even if the reforms are officially adopted as policy and embraced by the higher-level administration, the

[11] Recent theoretical work has shown how modifications to the design of RCTs can be implemented to enhance external validity of experiments when respondents select themselves into treatment (Chassang, Padró i Miquel, and Snowberg, 2012). See Berry, Fischer, and Guiteras (2015) for an application.

[12] Allcott (2015) also compares microfinance institutions that have partnered in recent RCTs with a global database of microfinance institutions and finds that partner institutions are older, larger, have portfolios with lower default risk compared with the broader population.

programs may fail because teachers never take them up. For example, the Coalation for Evidence-Based Policy reviewed 90 evaluations of educational interventions in the United States commissioned by the Institute of Educational Studies. 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 (Coalition for Evidence-Based Policy, 2013).[13]

Studies rarely document implementation challenges in great detail, but there some examples. Bold et al. (2015) replicate an intervention evaluated in Duflo, et al. (2011, 2015) where, in Kenya, an NGO gave grants to primary school parent-teacher associations to hire extra teachers in order to reduce class sizes from their very large initial levels. There were two versions of the program: an NGO-run version, which produced very similar results to the Duflo, et al. (2011, 2015) evaluation, and a government-run version, which did not produce significant gains. Analysis of process data finds that government implementation of the program was substantially weaker than NGO-led implementation: the government was less successful in hiring teachers, monitored the teachers less closely, and was more likely to delay salary payments. In addition to these implementation challenges, 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.

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 NGO-led intervention in India (Banerjee et al. 2007), the authors find negligible effects of the program on learning. They attribute this finding to a failure of teachers to integrate computer-assisted-learning into the curriculum. Banerjee, Duflo, and Glennerster (2008) report the results of an experiment that

[13] Interestingly these interventions were themselves often quite small scale, despite being scale ups of other even smaller studies.

evaluated a scheme that provided incentives for verified attendance in government health clinics in India. Although a similar incentive scheme had previously been proven to be effective when implemented in NGO-run education centers 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.

Several other studies have found that government implementation of programs can be incomplete. Banerjee et al. (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 the evaluation of the electronic payment system in India referenced above (Muralidharan et al., 2015) only about half of transactions in treatment areas were being made electronically after two years.

An interesting counter-example is offered by Banerjee, Hanna, et al. (2016) who 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. Interestingly, as also reported in that paper, 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 suggesting that a part of the observed initial failure of government implementation may have been due to its pilot status—in the full implementation, the government can use its regular delivery mechanism that are not necessarily activated in a pilot. As this discussion should have made clear, 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; the issue of equilibrium is addressed by large-

scale experiments (discussed in this issue in the paper by Muralidharan). The issue of "loss in transmission" is 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 this section, we illustrate how all these issues play out in practice by describing the long journey from the original concept of a specific teaching intervention, through its multiple variants, to the eventual design and evaluation of two "large-scale" successful incarnations implemented in government schools which are now in the process of being scaled up in other government systems.

3 An Example of a Successful Scale-up: Teaching at the Right Level

Our example of a program that eventually got scaled up is a remedial education program designed by the Indian NGO Pratham. The deceptively simple original idea behind these interventions is what we now call "teaching at the right level" (TaRL).[14] In many developing countries, like India, teachers are expected to teach a very demanding curriculum, regardless of the level of preparation of the children. As a result, children who get lost in early grades never catch up (Muralidharan 2016). Pratham's idea was to group children according to what they know (by splitting the class, organizing supplemental sessions, or re-organizing children by level) and teach them at the level they are at.

3.1 From Bombay Slums to 33 million children: scaling up without the government

The partnership between the researchers and Pratham started with a "proof of concept" RCT of Pratham's "Balsakhi" Program in the cities of Vadodara and Mumbai, conducted in 2001-

[14] Pratham credits Professor Jalaluddin, a known literacy expert, for developing the first incarnation of the pedagogy (Banerji, Chavan, and Rane 2004). The Pratham approach is called "teaching at the right level" and is also referred to as CAMaL – Combined Activities for Maximized Learning.

2004 (Banerjee et al. 2007). In this program, 3rd and 4th 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 community members (balsakhis) hired and trained by Pratham. Their learning levels (measured by 2nd grade-level tests of basic math and literacy) increased by 0.28 standard deviations.

After the balsakhi program, and partly propelled by its success, Pratham entered a phase of particularly rapid expansion. It took its approach out of the context of the relatively prosperous urban centers in West India into rural areas everywhere and in particular into the more "backward" Northern India.[15] As Pratham increased the scale of its program, the key principle of teaching children at the appropriate level remained, but it changed one core feature of its model to remain financially sustainable. Rather than paid teachers, it decided to rely largely on volunteers; these volunteers worked outside the school running their own learning-improvement classes and were much less closely supervised after the initial two weeks training. To facilitate this, the pedagogy became more structured and more formal, with an emphasis on frequent testing. Whether the model of remedial education could survive the 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.

The second RCT was conducted in Jaunpur district of Uttar Pradesh in 2005-2006: this was a test of the volunteer-led, camp-based Learning-to-Read model, in a rural area. The results were very positive: focusing on the treatment on the treated effect for children who participated, attending the classes made children 22.3 percentage points more likely to read letters and 23.2 percentage points more likely to read words. Nearly all the children who attended the camp

[15] By 2004, Pratham worked in 30 cities and 9 rural districts. (Banerji, Chavan, and Rane 2004)

advanced one level (e.g. from reading nothing to reading letters, or from reading words to reading a paragraph, etc.) over the course of that academic year (Banerjee, et al., 2010). This second study established that the pedagogical idea behind the balsakhi program could survive the change in context and program design.

Nonetheless, the study's process evaluation revealed new challenges. Because it was volunteer-based, there was substantial attrition of volunteers, and many classes ended prematurely. And since the program targeted children outside of school, take-up was far from universal. Only 17% of eligible students were treated. And most concerning, the program was not effective in reaching students at the bottom end of the distribution—those who were unable to recognize letters or numbers.

Thus, Pratham managed to increase its coverage on the extensive margin—targeting rural kids in new regions of India, but found a reduction in coverage at the intensive margin, particularly for those most in need. Nevertheless, in 2007, building on the success of the Learning-to-Read intervention and helped by the rigorous evidence demonstrating its effectiveness, Pratham rolled out its flagship "Read India" Program. Within two years, the program reached over 33 million children. However, this was only a fraction of school-age children in India, and did not appear sufficient to make a dent in the national statistics: Pratham's own annual nationally-representative survey (the Annual Status of Education Report or ASER) found no evidence of improving test scores.

To reach all of the children who needed remedial education, Pratham decided to revisit its scale-up approach to reach children while in school. To achieve this, they started working in partnership with governments. By 2009, a number of state governments were already collaborating with Pratham in running the Read India Programs. But whether the government's

implementation of the program was working was again an open question.

3.2 A first attempt to scale-up with government: the challenges

Starting 2008, J-PAL and Pratham 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 2008-09 and 2009-10. Importantly, 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 in Bihar was being run across 28 districts, 405 blocks, and approximately 40,000 schools, thus reaching at least 2 Million children. In Uttarakhand, the June before the evaluations were launched, Pratham was working in all of 12,150 schools in 95 Blocks of the State (Kapur and Icaza, 2010).

In addition to the RCT, we collected extensive process data, and partnered with some political scientists who, through interviews, collected invaluable details of the relationship between Pratham and the government. A companion working paper, Banerjee, Banerji, et al. (2016) provides more details on the evaluation design and the result of these two experiments as well as the two 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; and Sharma and Deshpande (2010) is a qualitative study based on interviews with parents, teachers and immediate supervisors of the teachers.

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 school teachers. Pratham provided materials and training for government school teachers, and also trained volunteers who supported teachers in the classroom. The government school teachers 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 (referred to hereafter as the "M" treatment). The second variant of the intervention included materials, as well as training of teachers in Pratham methodology and monitoring by Pratham staff (referred to as the "TM" treatment). 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 "TMV" treatment). In Bihar, the volunteer part of the TMV intervention was a replication of the successful Read India model evaluated in Jaunpur, since the volunteer conducted evening learning camps, focusing on remedial instructions (teachers were involved in that they were the one directing the children to the volunteer). Uttarakhand, on the other hand, had just the TM and TMV treatments and in the latter, volunteers worked in schools and were meant to support the teachers. In both states, 40 villages were randomly assigned to each treatment group.

The results (shown in table 1) were striking and mostly disappointing. The introduction of Pratham's methodology in schools during the school year, failed in both states. The M and TM in Bihar and the TM and even the TMV in Uttarakhand had no discernible impact. However, the TMV results in Bihar suggest that this was not simply context dependence: there we found a significant impact on reading and math scores, quite comparable to the earlier Jaunpur results. Since the TM piece seemed to make no difference, this suggests that the TMV intervention was just a volunteer intervention like that in Jaunpur. The pedagogical approach worked in this new context when implemented by volunteers. But the teachers were not able to implement the teaching-learning approach in the way that it was originally designed.

This is similar to the results of Bold et al (2015) in Kenya, who were able to replicate the

original results when the program was run by an NGO, but not when it was run by the government. The summer camp results, however, provided some hope. In just a few weeks of summer camp, there were significant gains in language and Math. The treatment on the treated effects were of the order of 0.4 standard deviations. This suggests that teachers were in fact able to deliver remedial education if they did focus on it, and that the failure of the model came from the fact that the more comprehensive teaching at the right level was not actually put in practice during the school year.

The process data and the qualitative information bolster this interpretation. Table 2 (panels A and B) shows some selected process measures. The situations were very different in the two states (see 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 (two thirds of the 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 also infighting within the educational bureaucracy, and strikes by teachers and their supervisors (unrelated to the program). Pratham staff was also demoralized and turned over rapidly. As a result, only between 28% and 45% of teachers got trained (for only three days each), and only a third of the schools used the materials, which they got very late. In many cases, there was either no volunteer or no teacher in the school during the monitoring visits.

What is common to the two states from the process data, however, is that a key component of Pratham's approach, the focus on teaching at the children's level were generally not implemented in schools, even in Bihar. Only between 0% and 10% of the classes in Bihar were observed to be grouped by levels. One consistent lesson of those 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 that was their express purpose. The biggest challenge for Pratham was how to successfully 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 (2011), Banerjee and Duflo discuss this resistance and point out the fact that TaRL is not being implemented in private schools despite the fact that most children in private schools are also not at the grade level. Since private schools are subject to lots of competition and do not lack incentives, they propose the hypothesis that teachers and parents must put much more weight on covering the grade-level curriculum than on making sure that everyone has strong basic skills. This is consistent with what qualitative studies reveal: teachers in both states seem to believe 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 "the 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 is correct it suggests two main strategies. Either to convince the teachers to take TaRL more seriously by working with their superiors to build it into their mission; or to 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 evening or summer classes. These ideas guided the design of the next two interventions.

3.3 Designing successful interventions

3.3.1 Working with Teachers: Getting Teachers to Take the Intervention Seriously

In 2012-13, Pratham, in partnership with the Haryana State Department of Education, adopted new strategies to embed the TaRL 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". To promote organizational buy-in, Pratham and Haryana tried incorporating TaRL into formal government training, formal systems of monitoring, and make the school time allocated to TaRL explicit and official.

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. While it had been the case in Bihar, teachers did not perceive it this way, in part because it was not relayed by their immediate supervisors, Cluster Resource Centre Coordinators. These coordinators had been invited to the training, but on a voluntary basis, and the responsibility of monitoring the teachers was left to Pratham staff. They never felt engaged with the program or felt that they were accountable for its success. In Haryana, to make the buy in by school system evident, one important innovation was the creation of a system of academic leaders within the government that could guide and supervise teachers as they implemented the Pratham methodology. As part of the interventions, Pratham gave the Associate Block Resources Coordinators (ABRCs—equivalent to Bihar's Cluster Resource Centre Coordinators) four days of training and field practice. ABRCs were then placed in groups of three in actual schools for a period of 15-20 days

to carry out their own daily classes and "test" the Pratham methodology of grouping by level and of providing level-appropriate instruction. Once the "practice" period was over, ABRCs, assisted by Pratham staff, in turn trained the teachers that were in their jurisdiction.

The second important feature is that the program was implemented during a dedicated hour during the school day. Beginning in the 2011-12 school year, the Government of Haryana mandated that all schools add an extra hour of instruction to the school day, for all schools. In regular schools, the normal school day was just longer. Within TaRL schools, the extra hour was to be used for class reorganization and teaching remedial Hindi classes using the Pratham curriculum. Reserving the same specific hour for restructuring the classroom across all schools 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 TaRL schools, all children in grades 3-5 were reassigned to ability-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 ABRCs. Once classes were restructured into these level-based groups, teachers were allocated to the groups for instruction. This made the grouping by ability automatic. This new version of the program was evaluated in the school year 2012-2013 in 400 schools, out of which 200 were selected to receive the program. The results, shown in table 3 (panel A) were this time positive: Hindi test scores increased by 0.15 standard deviations (significant at the 1 percent level). In this case, the intervention did not target math (there was no material or teacher training for math), and we find no effect there.

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

surprise visits. 95% of teachers in the treatment group, and virtually no teachers in the control group attended training. Most importantly, grouping by ability was also successful in Haryana, where it had largely failed in Bihar and Uttarakand: Over 90% of schools were grouped by learning levels during the time reserved for TaRL. In addition, teachers in Haryana used Pratham materials in 74% of the classes reserved for TaRL, where much lower rates were observed during the interventions in Bihar and Uttarakand. Interviews with teachers and headmasters and department administration suggested that the monitoring and mentoring role played by ABRCs was critical. Research staff helped set up a monitoring system and taught the ABRCs how to monitor teaching activities, with the intention of monthly visits to ensure schools were implementing the TaRL treatment. Indeed, 80% of schools reported a visit from an ABRC in the previous 30 days. Of those who reported a visit, 77% said that the ABRC spent over an hour in the school, and 93% said that the ABRCs observed a class in progress during at least one visit.

3.3.2 Using the schools but not the teachers: In-School Learning Camps

The alternative strategy, as we note above, was to use an outside team that came to the school during school hours. The underlying insight was that in areas where the teaching culture is very weak it is perhaps too costly to try to involve the teachers in this alternative pedagogy. It may make sense to use an outside team to sidestep the teachers and still take advantage of the school infrastructure and the fact that the children do come to school. The danger in going down this path, as we had seen in Uttarakhand before, was that the volunteers would be absorbed by the system, and end up working as substitute for the teachers.

To address this, 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. Pratham team members lead the teaching and are assisted by local village volunteers, and Pratham staff also regularly monitor the camps in each school and assist the school level team of Pratham and volunteers in administering the camps. Confined to short bursts of a 10 or 20 days each (and total of 50 days a year), they were more similar to the original volunteer ("learning to read") Read India 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, solving the previous problem of low enrollment. 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 again 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, 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. Panel B of Table 3 displays the impacts of the Uttar Pradesh interventions. The two interventions had similar impacts, with test score gains of 0.6 to 0.7 standard deviations. It is useful to pause and reflect how large these effects are. Figures 1 and 2 summarize visually the results in Haryana and Uttar Pradesh. The treatment effect is so large that by endline, treated children entirely catch up with the Haryana control children, and almost reach the level of the treated children in Haryana (in Uttar Pradesh, 48% of the treated children can read at the grade 2 level at endline; in Haryana, 47.5% of the control children can, and 53% of the treatment children), despite a huge gap (20 percentage point difference) at the baseline. This reflects in part the abysmal

performance of the school system in Uttar Pradesh, where very little is happening in control group schools: the number of students who cannot recognize any letter between baseline and endline in the control group fell from 34% to 24% in Uttar Pradesh, while it fell from 27% to 8% in Haryana. The number of students who can read at grade 2 level increased from 14% to 24% in Uttar Pradesh, compared with 34% to 47% in Haryana. 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 was set-up to collect data on attendance, evidence of learning by "grouping", activities during "camp" sessions, teaching practices of volunteers, involvement of school teachers and their perception of "camp" activities. There was strong adherence to key program components in Uttar Pradesh. 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, over 80 percent of classes in both treatments were observed to be grouped by achievement. There are four main policy lessons from this series of experiments with Pratham. First, there is clear evidence that the pedagogy that Pratham developed can improve basic learning levels in both reading and math. Second, this method can be effectively implemented even by village-level volunteers without formal teacher training, and by existing government teachers after relatively short-duration trainings on how to do this. Third, our process and qualitative data show that aligning teaching and materials to the child's initial learning level is key to successful implementation. However, fourth, achieving the alignment between pedagogy and initial learning levels requires an explicit organizational effort to ensure that children are assessed, grouped and actually taught at the right level; this will not occur automatically within the existing government school system, but can be achieved by articulating a narrow set of

program goals, ensuring there is specific time for the program, and properly supervising implementation. It took five RCTs and several years, to go 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, reaching almost 5 million children. The in–school volunteer led model has been implemented in 4210 schools across India, reaching over 200,000 children.

4 General Lessons

In this section we provide general lessons from the series of experiments described in the previous section, and a description of the "R&D" process that goes from formulating policy ideas to developing scalable policies.

Formulation of a successful policy begins with the identification of a promising concept. Small-scale experiments with NGOs can identify these concepts through both pinpointing the sources of specific problems and testing approaches of dealing with them. Fully understanding the key mechanism behind successful interventions may take more than one experiment. In the case of education, early experiments by Glewwe, Kremer, and Moulin (2009) and the initial Balsakhi results (Banerjee et al., 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). Inspired by 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.

The next step is to turn the insight into a program. In principle there are two separate steps here. First, a proof of concept study where we test the policy under more ideal conditions of implementation. This is where failures can be particularly informative since it is unlikely that the lack of results was because of the way it was implemented. There is probably no reason to go to the scale up stage in such cases.

Next, once a program has been chosen to be scaled, we need to develop an implementable large-scale policy version of it. This requires combining a good understanding of the mechanism underlying the concept with insight into how the particular government (or other large-scale implementer) works as an organization, to design the intervention such that it can indeed "stick". However even when this is competently done, there is no way to be sure that it will work; the only way to know what would happen when the implementation is done by a government on a large scale is to try it at that scale, to give a chance for all the potential problems that we have discussed to arise. And the only way to know if these problems have arisen is to do an RCT (or perhaps a high fidelity quasi-experiment) of the scaled program.

It is worth emphasizing however that the RCT of the scaled program does not have to be at scale; the sample needs to be large enough to answer the questions asked of it, and possibly diverse enough to be generalizable. It does not need to be at the scale of the intervention. For example the Bihar Read India study happened during a 28 district scale up of the program but covered only a sliver of it.

It is also not enough to document just the impact; descriptive data and process evaluations play an important role in helping us formulate hypotheses about why one thing worked and another did not and to be able to feed that into the design in the next experiment. Without the process evaluation and the two qualitative observation reports, we would not have had enough

information to develop the effective versions that were tested afterwards. These are ambitious projects, which take more time and resources, and there are relatively few examples to date, which makes it valuable to document them. Of the potential scale up issues we identified, which were the ones that turned out to be relevant in the Pratham/TaRL example?

Equilibrium effects did not arise in this context, despite the size of the scale up in which the evaluations were embedded. When well implemented, the program had large effects, even on large scale. Children are taught in their school, not pulled away, and we are not concerned about the returns to literacy falling down, as it seems a worthwhile objective.

The interventions were repeatedly stress-tested for *context dependence* by moving the program from urban India to rural UP, and then to Bihar, Uttarakhand, Haryana, and back to UP again. Moreover, complementary data emerged from Ghana, where a successful replication of the TaRL approach was organized with the government (Innovations for Poverty Action, 2015), and Kenya with the tracking program (Duflo, Dupas, Kremer, 2011). The ASER results, and results from similar tests worldwide, made it clear that there are many children who clearly needed remedial education.[16]

Although there were political issues 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. This does not mean they would not exist elsewhere. An attempt to pilot the project in another state, Tamil Nadu, was not successful after the government officials displayed strong resistance to working with Pratham, which had exposed the less than stellar performance of government schools with ASER: Pratham has become such a large player

[16] In terms of understanding the need to remedial education it is striking that the effect of the volunteer-run program run in schools in UP are as high in *reduced form* effect as the instrumental results were in the first Read India evaluation in UP: this suggest that the high early results were not driven by compliers with very high marginal returns in the original experiment.

in the India educational scene that they are not just a regular partner to work with. Tamil Nadu government had their own pedagogical approach called "Activity Based Learning" which they were not keen to subject to scrutiny.

The key obstacle in this case was the difficulty of implementing at scale. The first successes were clearly affected by a combination of *randomization bias* and *piloting bias.* Pratham was one of the first organizations to partner with researchers to evaluate its programs (before J-PAL even existed), and may be somewhat unique in its combination of scale and purpose. It is conceivable that moving to any other partner (not just the government) would have been difficult. We do not know this for sure since we did not attempt replicating with another NGO in India, but even within Pratham, according to the institutional analysis, it was harder to find a good team in Uttarakhand than in Bihar, where there was much more general enthusiasm. 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 TaRL

The process of policy design is necessarily iterative. In our case, while a successful concept was identified after the 2005-2006 Uttar Pradesh study, it took 4 subsequent RCTs over a number of years to go from a concept to a policy that actually could be successful on a large scale in government schools. During this entire period, Pratham continued to evolve its programs and implement on scale. These ongoing interventions on scale provided a good stage on which to periodically launch impact evaluations. The good working relationship of Pratham with JPAL made it conducive to conduct the sequence of RCTs over time. Each experiment built on past successes as well as failures, with modifications to both the design of the intervention and in the manner of implementation within and across experiments. In the end, two successful scalable interventions have been shown to be effective, but these interventions would not have been

identified without learning from the failures along the way. This process of learning could be formalized, and the behavior of government bureaucracy as organizations could themselves be the object of the research (and possibly experimentation) in the process of scale up. This would be a fascinating research program on its own right.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics, Forthcoming*.

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." *NBER Working Paper*, no. 15794.

Athey, Susan, and Guido Imbens. 2016. "The Econometrics of Randomized Experiments." In *Handbook of Field Experiments*.

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: 37–66.

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."

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, 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, Sylvain Chassang, and Erik Snowberg. 2016. "Decision Theoretic Approaches to Experiment Design and External Validity." In *Handbook of Field Experiments*.

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." *Mimeo, Yale University*.

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, and Esther Duflo. 2011. *Poor Economics*. PublicAffairs.

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

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, Clement Imbert, Santhosh Mathew, and Rohini Pande. 2016. “E-Governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India.” *Mimeo, MIT*.

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.” *Mimeo, MIT*.

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*. http://www.india-seminar.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 Controlled Trial of a Language Arts Program.”

Berk, Richard A, and Lawrence W Sherman. 1984. “The Minneapolis Domestic Violence Experiment.” *Police Foundation*.

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 No. DP10703*.

Blair, Graeme, Radha K. Iyengar, and Jacob N. Shapiro. 2013. “Where Policy Experiments Are Conducted in Economics and Political Science: The Missing Autocracies.”

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.” *Working Paper*.

Brigham, Matthew R., Michael G. Findley, William T. Matthias, Chase M. Petrey, and Daniel L. Nielson. 2013. “Aversion to Learning in Development? A Global Field Experiment on Microfinance Institutions.” *Annual Convention of the International Studies Association*.

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*, no. 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.”

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. 2013. “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. doi:10.1257/jel.48.2.424.

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 (August): 1739–74.

———. 2015. “School Governance, Teacher Incentives and Pupil-Teacher Ratios.” *Journal of Public Economics* 123: 92–110.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. “Incentives Work: Getting Teachers to Come to School.” *American Economic Review* 102 (4): 1241–1278.

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, and Edward Miguel. 2016. “Impacts and Determinants of Health Levels in Low-Income Countries.” In *Forthcoming, Handbook of Field Experiments*.

Evidence Action. 2016. “Kenya Deworming Results Announced: 6.4 Million Children Worm-Free and Ready to Learn.” Accessed August 27. http://www.evidenceaction.org/blog-full/kenya-deworming-results-announced-6-4-million-children-worm-free-and-ready-to-learn.

Glennerster, Rachel. 2016. “The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency.” In *Handbook of Field Experiments*.

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.

Hausman, Jerry A, and David A Wise. 1985. *Social Experimentation*. National Bureau of Economic Research Conference Report. http://press.uchicago.edu/ucp/books/book/chicago/S/bo3625566.html.

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

Heckman, James J., and Edward J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments." In *Handbook of Econometrics*, 6B:4875–5143. Elsevier B.V.

Innovations for Poverty Action. 2015. "Targeted Lessons to Improve Basic Skills." Policy Brief.

Kapur, Avani, and Lorenza Icaza. 2010. "An Institutional Study of Read India in Bihar and Uttarakhand." *Mimeo, J-PAL*.

Kremer, Michael, and Rachel Glennerster. 2012. "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. Vol. 2. Elsevier.

Lengeler, Christian. 2004. "Insecticide-Treated Bed Nets and Curtains for Preventing Malaria." *Cochrane Database of Systematic Reviews*, no. 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.

Meager, Rachael. 2016. "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments." *Mimeo, MIT*.

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

Muralidharan, Karthik. 2016. "Field Experiments in Education in Developing Countries." In *Handbook of Field Experiments*.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2015. "Building State Capacity:

Evidence from Biometric Smartcards in India." *Mimeo, University of California San Diego*.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics, Forthcoming*.

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 Papers and Proceedings* 105 (5): 471–75.

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

Sandefur, Justin. 2011. "Held Hostage: Funding for a Proven Success in Global Development on Hold in Kenya." *Center for Global Development*. April 25. http://www.cgdev.org/blog/held-hostage-funding-proven-success-global-development-hold-kenya.

Sharma, Paribhasha, and Anupama Deshpande. 2010. "Teachers' Perception of Primary Education and Mothers' Aspirations for Their Children - A Qualitative Study in Bihar and Uttarakhand." *J-PAL South Asia*.

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*, no. 7.

Vivalt, Eva. 2015. "How Much Can We Generalize from Impact Evaluations?" *Mimeo, New York University*.