Center for Global Development

CGD NOTE
DEC 2018

It's Like That and That's the Way It Is?

Evaluating Education Policy

Susannah Hares

It's tricky to evaluate government education policies. They're not implemented in NGO-like laboratory conditions, and political motivation and public sector capacity constraints play as much of a role in their success or failure as policy design. Using the examples of three rigorous studies of three different education policies, this note aims to shed some light from the perspective of someone on the policy side on how, why, and when to evaluate government-led reforms.

A government education policy is not an abstract theory that can easily be replicated in a different place. In each new context, it is effectively a brand-new programme and needs to be evaluated as that. None of the three examples presented was "new" as a policy: school inspections, school vouchers, and charter schools have all been tried and evaluated elsewhere. But the evaluations of these policies—when implemented in new contexts—illuminated a new set of challenges and lessons and generated a different set of results.

Evaluation needs to be part of the process of designing, iterating, and implementing education reforms—it shouldn't be used to reach a conclusion about the universal effectiveness of any one policy. Good evaluators work together with government partners, embracing each unique context and its complexity, rather than conducting detached research in the pursuit of a single truth.

It's hard to make good education policy, especially in places with limited resources and fragile institutions. Theoretically at least, doing something new almost always means not doing something else. If more teachers are trained, perhaps fewer textbooks are bought. If more money goes into secondary schools, it might be at the expense of primary education. Budget trade-offs are real.

Given the huge need and the scarce resources available, education policymakers, policy advisers, and donors *should* want independent evidence telling them (a) if a policy is delivering better outcomes at an affordable cost and (b) whether the policy is better than viable alternatives. And they *should* want to act on the evidence generated to expand, course-correct, or stop the reform. But education reform usually takes place in a highly charged political context and the reality is far more complex. A low or precisely estimated zero effect in a study does not always mean a policy will—or indeed should—be thrown out. And likewise, a very positive effect doesn't mean it will or should be scaled nationally.

Ark—whose global education policy programme I led for seven years—regularly faced these issues in its work advising ministers of education on education reform. The global education reform movement—fondly referred to by its critics as the GERM—is controversial. The projects Ark worked

on were often the subject of heated domestic and global debate. Cognisant of this, we felt it was all the more important to be generating truly objective evidence on the impact of our work so that we, our government partners, donors and other observers knew which projects were working, when they needed improving, and when they had "failed."

During my time at Ark, we helped set up independent, randomised control trials (RCTs) on three different—but all big and ambitious—education reforms. The aim was to generate the strongest possible evidence of causality. The mere commissioning of these evaluations was oftentimes controversial. A few folks said a policy shouldn't be evaluated until it's been iterated and tweaked so as the design is just right, the programme has been expanded, and any wrinkles have been ironed out (although by this point, the policy may well be too big or too political to fail). Some argued that policies "proven" elsewhere don't need evaluating again—more expensive studies are a waste of time and money. Others suggested that independent evaluations are a form of accountability—an unhelpful pass/fail judgement on performance rather than an opportunity for policymakers to get good data to make informed decisions.

On the surface some of these critiques make sense and the Ark team considered them all, together with our research and government partners. But ultimately, the three examples presented show that the context of the education reform, and the timing and independence of evaluations, really do matter.

"FAILURE IS THE KEY TO SUCCESS; EACH MISTAKE TEACHES US SOMETHING." — Morihei Ueshiba

In his book *Failing in the Field* Dean Karlan writes about why project fail, drawing inspiration from Ueshiba's quote. His thesis is that we can and should learn as much from what didn't work as we do from what did. From design flaws, through implementing partner challenges, to inappropriate settings, he presents a set of case studies that provide a series of lessons for practitioners and researchers undertaking evaluations of NGO projects. The lessons are insightful. Indeed David McKenzie, in his review of the book, notes that the main causes of failure of those NGO projects—with the gift of hindsight—could probably be controlled and prevented.

The projects presented in this note are different in that they are evaluations of government-led education policies. Working with governments brings its own set of considerations: capacity issues within bureaucracy; lack of control over programme design; political motivations. The reasons why one might undertake an evaluation—and when—are quite different. Politicians generally want to see things scale quickly, so there may not be time or space to tinker with policy design in a pilot setting before rolling it out. The reforms often take place in a highly charged political environment, with evidence of impact only being one—sometimes minor—consideration for decisionmakers.

In contexts like this, without independence "failure" may not be detected or acknowledged. But a good independent evaluation isn't about pass or fail—it's much more nuanced than that. It provides a lot of really useful insights about various dimensions of the policy, which *should* aid reflective conversations with policymakers and help governments with the "what next" decisions.

It's fair to say none of the policies presented scored anywhere near a perfect 10; I can't think of any education policy, anywhere, that does. But neither were any full-on failures. The studies all generated robust evidence and valuable insights that helped Ark, its government partners, and the sector broadly learn more about the viability of various education policies—school inspections, school vouchers, charter schools—in different contexts. Indeed, the studies have *all* provided important lessons that

Ark has taken to other contexts, whether that's charter school policy design in Ghana, private school subsidies in Uganda, or school inspections in South Africa.

Taken together, the evaluations of the three projects presented shed some light on how we should and shouldn't be evaluating big and ambitious policy reforms.

INSPECTING SCHOOLS IN MADHYA PRADESH

Back in 2012, the government of Madhya Pradesh—home to 112,000 public schools—asked Ark to help design a school quality assurance programme. The programme draws from global best practice and aims to both measure school quality *and* track school improvement. Schools conduct a detailed self-assessment—validated by external assessors—and a targeted school development plan is subsequently produced identifying areas of improvement. Progress is then supposed to be tracked over time through twice-yearly school-level follow-ups by district education officers. The programme started with two years of design work and a pilot of 100 schools, and was then expanded rapidly to 25,000 schools, with strong support from the highest echelons of state government.

To evaluate the policy, the research team randomly assigned 2,000 schools in five districts to the treatment group and effects were studied 18 months later. They looked at implementation fidelity, student outcomes (measured through independently administered assessments), and observed measures of governance and pedagogy.

Results seemed promising at first. There was near-universal implementation of the school assessments—around 90 percent of schools had undertaken the assessment and prepared school improvement plans. Sadly, that's where the good news ended. There was no change in the functioning of the schools—no extra monitoring or accountability; no improved classroom practices. And there was no improvement in test scores. Nothing. After six years of hard work, this was tough for the government and the Ark team to hear!

But we were glad we heard it. The study provides great insights for this project and—given its scale and rigour—for accountability initiatives elsewhere in the world. It contains important lessons about top-down policymaking and the huge gulf between policy design (which in Madhya Pradesh was pretty good), and execution (which was poor). Had the team not undertaken the study, Ark and the government might well have continued on the same track, with significant investments of time and resources. While the null result on the 2,000 schools did not directly influence the government's decision to scale the project to 25,000 schools, there was certainly deep engagement with the "failure" on the part of the government, and iterations on the policy design were implemented during the scale up to address the issues identified by the evaluation.

The zero-effect result doesn't say "accountability policies don't work anywhere." Instead, it suggests policy advisers should consider how these reforms should be implemented given the execution capacity constraints of a particular system.

GIVING KIDS VOUCHERS FOR PRIVATE SCHOOLS IN DELHI

In 2009 the government of India launched the Right to Education Act (RTE), an intriguing policy that mandates all private schools to reserve 25 percent of their places for kids from disadvantaged families. It aims to give poor children an opportunity to attend better schools, with an inbuilt assumption—shared by many at the time—that private schools perform better than public schools.

To explore whether this policy had the potential to deliver on its objectives, Ark ran a school voucher lottery that provided households with tuition-fee free access to low-cost private schools in Delhi. Treated students were given five years' worth of tuition vouchers to attend a nearby private schools. After six years, the research team tracked the voucher winners and losers, collecting information about their academic abilities, their schools, and their aspirations. They found that winning the voucher had no impact on maths or English, and actually lowered Hindi scores. No effect was found on noncognitive skills, parent aspirations, or social networks.

We weren't the only ones researching various dimensions of this policy. A study by Vijay Kumar in Karnataka found no effect on learning outcomes, although he did find a positive effect on self-efficacy. Gautam Rao also found no effect on test scores, but his study showed that rich kids who shared classrooms with poor kids were a bit nicer. A seminal study by Muralidharan and Sundararaman in Andhra Pradesh found no significant effect on Telugu, maths, English, and science tests scores, but a strong positive effect on Hindi scores (which is taught in private but not public schools). And the same patterns were evident in panel data from Andhra Pradesh presented in a paper by Abhijeet Singh.

These studies could lead you to conclude that the policy will not deliver on its intended objective of improving learning for poor kids. However, RTE is possibly the largest public-private partnership (PPP) in the world, and the government of India probably won't abandon it any time soon. There may well be ways to iterate the design and target the policy more effectively so as to better achieve its aim. Results from studies like these should encourage policy advisers and policymakers to scrutinise policy design and execution, as well as any plans to scale it. Simply ceasing a policy after a disappointing result could mean throwing the baby out with the bathwater.

OUTSOURCING PUBLIC SCHOOLS IN LIBERIA

Liberia's education system is among the most fragile in the world. Not only are learning levels dismally low, but access to basic education remains inadequate. With the system in "crisis," the Minister of Education looked for radical options to improve learning. Starting in 2016, with Ark acting as policy adviser to the ministry, 93 public primary schools were contracted out to eight private providers in a PPP called "Partnership Schools for Liberia" (PSL). PSL is a very public PPP—teachers remain on the government payroll, schools are free to attend and remain the property of government, and academic or other selection is prohibited.

Nevertheless, it was perceived as privatization and so the policy was controversial. PSL quickly became the cause of much debate within and outside of Liberia. On Ark's advice, the ministry decided to commission a rigorous, independent evaluation in year one, so as to provide themselves and the project's many partners and observers with objective and robust information on project and provider performance. This was a brave decision by the Ministry of Education, particularly when many of those partners and observers were drawing conclusions and running rival or self-commissioned studies about the project before the results were in.

A year later the results of the independent, government-commissioned study by Romero et al.'s research were actually in. Their release was quite a lesson in how different people can read the same report in very different ways. Some proclaimed the policy a huge success and insisted it should be scaled nationally with urgency. They said that to not do so would be the immoral withholding of a proven vaccine from Liberia's children. Others declared it a total failure and called for its immediate cessation.

As is usually true, the reality was somewhere in the middle. This was an evaluation of a multi-operator programme, and inevitably there was a range of performance. Overall, teachers in the outsourced schools were in school more and were teaching more. And learning levels increased by 60 percent—albeit from and to an unacceptably low level.

But costs were very high—between 2 and more than 20 times government budgets, depending on the provider. There were also serious unintended consequences: the largest provider—Bridge International Academies—unenrolled thousands of kids and dismissed more than half of the teachers in their schools. And recently, far more tragically, an outstanding piece of journalism investigating the behaviour of one of the PSL operators has shed light on the perils of letting private actors loose in fragile, low-capacity education systems like Liberia.

When it came to future financing and policy decisions, the independence and the rigour of the evaluation was crucial. With loud voices on either side urging for the immediate scale or termination of the programme, the study provided the government and donor partners with much-needed collateral to inform their decisions. And since the evaluation commenced at the start of the project, it provided decisionmakers with the information to differentiate—if they chose to—between the performance and the behavior of the eight private providers.

Beyond Liberia, the results raise questions about the viability of this kind of policy at scale, illuminating the risk of PPPs in low-capacity systems—a topic discussed in the 2018 World Development Report. Overseeing, monitoring, and regulating private contractors is challenging and expensive, and *may* not be any easier for a government than actually running schools. The evidence base on this question is limited and fiercely debated. The study by Romero et al. makes a valuable contribution to this global debate as well as to the domestic policymaking process in Liberia.

WHITHER EVALUATIONS OF GOVERNMENT EDUCATION REFORMS?

No one study can or should demonstrate once and for all the viability or otherwise of a particular education policy. The studies discussed in this note certainly don't suggest that the policies evaluated should be thrown out and never tried again. But they *do* suggest that it is absolutely worthwhile to generate evidence in a new context—even on a supposedly proven model—before trying to scale a particular policy.

This may seem inefficient. Some RCT-purists say that once something has been proven to work in one place, judgement has been passed and it should be scaled up there and elsewhere without further expensive evaluations. But the examples in this note show that the design of the policy or programme is only one—often minor—part of the equation. The implementation capacity or constraints of a particular system are significant drivers of success or failure. At the extreme, something that has worked in Finland is unlikely to work in Sierra Leone. Even between different states in India, we see very different effects from similar programs. And in Bangladesh, the RCT results of the scale up of "no lean season" were disappointing, after a very promising pilot programme. Given the scarce resources available, it seems sensible to establish what works and what doesn't work in a particular context before scaling.

The studies also reinforce the value of independent evaluation, especially in the frenzied, heated world of education reform. Many argue that rigorous impact evaluation should be about learning—researchers generating evidence to help implementers improve. If this is the case, perhaps the "independent" part of an independent evaluation is defunct. But independence *does not* mean

disengaged: a good research team, like those described in these examples, share insights and flag risks throughout the course of the project, rather than waiting until the results are ready to be published. They do not act as mere auditors, but—particularly for these large government reforms which have the potential to reach hundreds of thousands of children—as true partners in thinking through policy design, execution and outcomes.

What independence *does* mean is the collection of robust and objective data on education policy reforms, even if it doesn't always tell you what you want to hear (and making sure it is published, no matter what it says). It means not being beholden financially to those whose project is being evaluated. It means ensuring that the research design is guided by best practice, and not gamed to make individual operators look good or bad. Factors other than evidence drive policymaking in every part of the world: electoral incentives, donor fads and favourites, budget trade-offs, and so on. An independent study can be a vital tool to push back with when the pressure to scale fast is overwhelming. And an independent study allows the detection and discussion of "failure," whether or not decisionmakers want to act on that information.

Embracing precisely estimated zero- or low-effect RCTs—rather than carrying on regardless—can be painful, but it's necessary. I've often been struck by the amount of influence donors and advisers—usually outsiders—can wield over education policymaking in developing countries. With influence comes responsibility. RCTs and other rigorous evaluations can be expensive and difficult, and the results can be disappointing. But, if we're going to take big, ambitious shots at education reform, we need to invest in proper evaluations to generate truly objective evidence on their effect in a particular context. And we need have the courage to rise above the hype when the results are not what we want to hear.

With thanks to Mauricio Romero, Justin Sandefur, and Abhijeet Singh for their very helpful comments.

SUSANNAH HARES is a senior policy fellow at the Center for Global Development.

WWW.CGDEV.ORG