Experimentation, Innovation, and Economics

Prize Lecture, December 8, 2019 by
Michael Kremer
Harvard University, Cambridge, MA, USA.

The experimental method not only helps identify causal relationships, but also provides economists with a rich sense of context, focuses research on specific practical questions, stimulates collaboration with practitioners and specialists from other fields, and allows for rapid iteration. In this lecture, I present a series of examples illustrating how together these features make the experimental approach a powerful tool for advancing scientific understanding, informing policy, and promoting innovation. I then discuss how institutions can be designed to accelerate innovation and direct it toward the world's most pressing needs.

When I first began using the experimental approach in 1994, I saw it as a way to cleanly isolate causal impact from potential confounding factors. But it has since become apparent that field experiments create an opportunity to complement existing approaches in economics with a fundamentally new type of economics research.

This lecture first lays out some key features of the experimental approach. It then provides some concrete examples, illustrating how these features make field experiments an important tool for economic science, for policy, and for innovation. I will conclude by discussing the need not only for innovation itself, but for metainnovations: innovations in social institutions designed to accelerate scientific and technological change and orient it toward human needs.

I. KEY FEATURES OF THE EXPERIMENTAL APPROACH

When I began my career as an economist, I worked on questions relating to economic growth and technological change. Like most economists, I worked mostly alone or with one or two co-authors, using models and data which others had collected.

I believed at the time, and continue to believe now, that this approach to economics, broadly following the austere, modernist program laid out by Milton Friedman in "The Methodology of Positive Economics" (Friedman 1953), can be extremely valuable. But due largely to serendipitous factors, my research and much research in development economics has taken a different approach. I had taught secondary school in Kenya before going to graduate school, and in 1994, the summer after getting my first teaching job, I returned to Kenya to visit friends. One of those friends, Paul Lipeyah, had just been hired by an NGO focused on education. They wanted to know the impact of their work. Since they were deciding which schools to expand into first, I suggested that they could use random assignment to measure their impact, approximating a medical trial.

This work in Kenya certainly had precedents. Medical researchers have long used randomized trials, dating in part to the pioneering work of Phillip D'Arcy Hart and Austin Bradford Hill (Marshall *et al.* 1948) on tuberculosis clinical trials in the 1940s. Jamison's (Jamison et al. 1981) trial of radio mathematics education in Nicaragua was a pioneering use of trials in development economics, and the Rand Health Insurance experiment (Newhouse 1993), the Negative Income Tax experiment (Rossi and Rosenbaum 1983), and a series of job training experiments (LaLonde 1986; Hämäläinen, Uusitalo, and Vuori 2008) laid the groundwork for much future work.

I initially saw randomization and field experiments as a way to cleanly isolate the causal impact of a program or policy from potential confounding factors. But I have since come to realize that field experiments created an opportunity for a fundamentally different type of economics research that can complement other approaches in important ways.

The ability to isolate causal impact from confounding factors is a powerful tool. Field experiments have frequently shown that policies have very different effects than previously believed based on nonexperimental correlational evidence, both in developed countries (LaLonde 1986), and in developing countries (Muralidharan and Sundararaman 2015; Glewwe, Kremer, and Moulin 2009).

But in addition to isolating causal impact, field experiments have four other key features. Ironically, the very constraints imposed by experiments open up new opportunities by forcing us to work in new ways.

First, in part because they require time in the field, experiments provide economists a richer sense of context than many would otherwise obtain. While most empirical economists previously worked with existing datasets, the experimental approach requires researchers to talk to farmers, teachers, students, and small business owners where they live and work. Conducting focus groups, piloting questionnaires, and

becoming familiar with the details of the policies being evaluated are essential steps in the process of conducting an experiment. Thus, field experiments are useful not just for testing the researcher's pre-existing hypotheses, but also for generating new hypotheses.

Second, while experiments are often designed to shed light on larger conceptual questions, they typically also address very specific, practical problems. Focusing on specific, practical problems often allows us to see the importance of issues we had previously overlooked, and these often turn out to be of importance beyond the specific context.

Third, experiments are inherently collaborative, requiring us to work with practitioners in governments and civil society, teams of survey enumerators, and specialists in other fields such as education, health, agriculture, or psychology. This collaboration allows the ideas and experiences of a much broader set of people to enter economic research.

Fourth, the modern experimental approach in economics is iterative. One experiment might test the impacts of a multifaceted program. A second might seek to understand which components drive the effects, or test hypotheses regarding the underlying economic or psychological processes at play. A third might examine if the results could be improved with tweaks to program design, or if similar results hold in a somewhat different context. In some instances, a follow-up study could examine longrun impacts. In some cases, iteration may take place within a single research team, but in others it will be conducted by other research teams, as part of a decentralized process. The way we learn through the experimental approach is not a single breakthrough in one paper but the accumulated wisdom and insight from a series of studies. While centrally directed research programs have played a role, progress in the field has emerged primarily from a decentralized, selforganizing process in which many different researchers and implementers follow each other's work, each making their own decisions about the appropriate next steps.

Several features of the modern wave of experiments that began in the town of Busia in Western Kenya allowed for more rapid iteration than under the pioneering randomized controlled trials (RCTs) of the 1970s, which typically evaluated individual large scale government programs. The new wave of experiments were typically conducted by researchers working with NGOs or on their own, meaning researchers are involved not just in evaluation but in program design. The programs themselves are often financed out of research budgets, and so iteration does not have to wait until the government is prepared to pay for the roll-out or testing of a new national program. Relative to national governments, NGOs tend to be more nimble, are not expected or able to roll out services simultaneously over an entire country, and typically have programs that evolve over time. Over time, NGOs and their funders learned that they could increase their impact by innovating with new, potentiallyscalable approaches, working with researchers to carefully evaluate impact, and disseminating the results, so that other NGOs, firms, and governments can learn from their experience. Of course, the fact that research budgets and NGO budgets are much smaller than national government budgets meant that new approaches had to be developed to allow research to be conducted inexpensively, and that new institutions were needed to facilitate this.

Increasingly the two strands of work have merged. The seminal trial of conditional cash transfers by the Mexican government (Gertler 2004, Schultz 2004) was an evaluation of a government program, but it was a program designed by Santiago Levy, who had a research background. Subsequent evaluations tested program variants, often inspired by research. Researchers are increasingly working with governments on iterative evaluation of large-scale government programs (Barrera-Osorio *et al.* 2008, 2011; Muralidharan, Niehaus, and Sukhtankar 2016, 2017).

Together, these five characteristics make RCTs a valuable tool for economic science, for policy, and for innovation. Because Esther Duflo and Abhijit Banerjee provide compelling illustrations of the scientific and policy value of the experimental approach in their lectures, this lecture will emphasize the relationship between the experimental approach and innovation. Spending time in the field, collaborating with practitioners and specialists in other disciplines, studying specific, practical problems, and iterating within and across research teams lead naturally to innovation. The experimental method not only allows us to test new policy approaches but to delve into causal mechanisms, explore outcomes in different contexts, test variations, and iterate to produce improved solutions. I initially thought of experiments primarily as evaluation but now see many experiments as more akin to beta tests, useful in developing new products or policies and not just studying existing ones.

This innovation is useful not only in addressing immediate human needs, but also in advancing scientific understanding. In the mid-twentieth century, many thought of the relationship between scientific knowledge and practical change as unidirectional, with “pure” scientific research eventually leading to practical technological innovations. We start with Newton’s laws of physics, and then eventually get an airplane. First come the laws of thermodynamics, and then emerges the steam engine.

However, in fact, the relationship flows both ways. Key breakthroughs in scientific theory often result from pursuing new technologies. For example, the science of thermodynamics originated from the study of heat engines, including the steam engine, which were already in use (Erlichson 1999).

The experimental approach is built upon this twoway interchange of ideas between the applied and the abstract. It often starts with a real-world problem but yields general insights which are fed back into new policies and programs. The following examples demonstrate how the experimental approach can help in advancing economic science, informing policy, and sparking innovation.

II. EXAMPLES

A. Education

When I began conducting randomized evaluations in Western Kenya, a central debate in the literature on education economics was the relative importance of resources versus incentives. To overgeneralize a bit, some argued that providing more resources would improve learning (Card and Krueger 1992), while others argued that resources on their own would do little good and that teachers and administrators needed better incentives (Hanushek 1995). Even funding skeptics believed, based on common sense and correlational evidence, that remedying a shortage of textbooks would increase learning in a particularly cost-effective manner (Lockheed and Hanushek 1988). According to the Ministry of Education, on average in Kenya there was only 1 textbook for every 17 students. So, the NGO decided to provide textbooks, and Paul Glewwe, Sylvie Moulin, and I (2009) helped them evaluate the program, fully expecting that we would find that the causal impact was similar to or perhaps even larger than that suggested by the correlations. Instead, we could not reject the hypothesis that providing textbooks had no impact on average test scores.

This unexpected finding pointed us to a broader issue. Teaching in the schools we examined, as in many developing country schools, was oriented toward the highest performing students. At the end of primary school, only those students who could pass high stakes exams went on to secondary school. Among those, the very top performers qualified to attend more selective secondary schools. Schools were judged largely based on the scores of students at the top of the distribution and teachers focused on those students and endeavored to cover the full curriculum. But many other students were left behind. After the first few years of primary school, English is the medium of instruction in Kenyan schools, but this was the third language for most students in the schools we studied.

This suggested the hypothesis that if some students fell far enough behind that they could not follow the textbooks, providing textbooks would not do much good. We realized we could test this theory by looking at how providing textbooks affected students at different initial learning levels. Consistent with that theory, pupils with the highest baseline performance benefited substantially, but others did not (Glewwe, Kremer, and Moulin 2009). The combination of rigorous testing of causal impact and a strong sense of context enriched our thinking. We started off thinking about resources versus incentives, but the findings pointed toward a fundamental political economy issue: the orientation of education toward elites. Although this dynamic is now viewed as playing a major role in generating poor learning outcomes, it was not the focus of economists' thinking at the time.

Subsequent work has gone much further in pursuing this idea, producing both a deeper understanding of education in the developing world and policy solutions which have spread to reach millions of students. Studies conducted by Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden (2007) showed that targeting instruction to match a student's level of achievement could generate huge gains. With close involvement from Abhijit and Esther and other coauthors, organizations such as the nonprofit Pratham have widely deployed new educational approaches such as targeted instruction based on this body of work (see Banerji and Chavan 2019 for a history of the iterative process of developing Pratham's programs). These insights have been relevant beyond specific programs and modes of education delivery.

Karthik Muralidharan, Abhijeet Singh, and Alejandro Ganimian (2019) studied a computer-based program which automates and dynamically adjusts the targeting of instruction. This was found to be highly effective in improving learning across the distribution of initial test scores and is also being scaled up in India.

Grouping students together in classes by initial achievement rather than age is often seen as an approach that may disadvantage students with low initial learning levels. Esther Duflo, Pascaline Dupas, and I (2011) developed a simple model to show that in contexts in which teachers have incentive to focus on the top of the distribution, grouping students by initial achievement may make all students better off, and we found evidence that students across the distribution of initial test scores benefited from this approach in Kenya.

Together this series of studies suggests substantial political economy problems in education. While existing RCTs focus on strategies for ameliorating the problem, they provide a diagnosis that could inform wholesale reform efforts designed to structure the school system around the needs of the typical pupil rather than the children of the elite. Experiments reflect the range of policy options that are politically feasible, but should some countries decide to try more radical reforms, they could potentially be evaluated experimentally.

B. Worms, Pricing of Preventive Health Products, and Behavioral Development Economics

While providing textbooks had a smaller impact than we expected, another strategy – providing free deworming medicine for school children – led to large increases in deworming take-up and improved educational and economic outcomes. As discussed below, other experiments suggest this is part of a more general phenomenon in which even modest prices prevent people from investing in cost-effective preventive health goods, suggesting an important policy role for free provision. Many people seem to forgo high return investments more generally. Appropriately designed experiments can help tease out the reasons, and several point to the importance of nonstandard preferences, such as present bias and loss aversion, combined with naïveté regarding future preferences.

Worms

More than one billion people worldwide are at risk of intestinal worms such as hookworm, whipworm, roundworm, and schistosomiasis. Worms are particularly common among school-age children, which is why the World Health Organization recommends school-based mass treatment of children in endemic areas.

Edward Miguel and I (2004) used the staggered roll-out of an NGO schoolbased deworming program to examine its educational impact. In the short run, we found that the program reduced school absences by roughly one-quarter. It was not only the children who received deworming medication who benefited. Other children in the same school and even in nearby schools also had lower worm loads and better educational outcomes, presumably because mass treatment decreased disease transmission. We were able to collect follow-up data on participants in this program up to 20 years after the program and found large impacts. In the long run, the program increased the fraction of girls finishing primary school and entering secondary school – halving the gender gap in secondary school attendance – and increased students' future earnings enough to pay for the cost of the program many times over (Baird et al. 2016). In fact, just the increased tax revenues governments could collect as a result would more than offset program costs. While many early RCTs focused on short run effects, there now is a burgeoning literature using them to study long-run impacts (Bouguen et al. 2019).

With technical support from the Deworm the World Initiative and Evidence Action, the Kenyan government scaled the program nationally. Later, various Indian states and eventually the Indian national government adopted similar programs. More than 150 million children across Asia and Africa are now receiving deworming medicine each year.

Pricing of Preventive Health Products
When our work started in the 1990s, many policymakers and NGOs, including our partner NGO, thought it was important to charge people for preventive health products such as mosquito nets, water treatment solution, and deworming pills in order to promote financial sustainability, target the product to those who needed it most, and encourage people to use it. We convinced the NGO to instead try providing free deworming pills at a subset of schools. We found that even the NGO's small fee (<$0.40 per child) dramatically reduced take-up of deworming medicine to just 17 percent, compared to the 70 percent participation rate with no fee.

A series of experiments have found similar patterns in the response of adoption of preventative health products to pricing in other context (Dupas 2014; Cohen, Dupas, and Schaner 2015; Cohen and Dupas 2010) and cleverly tested – and found little empirical support for – hypothesized benefits of charging for preventive health

products (Dizon-Ross, Dupas, and Robinson 2017; Ashraf, Berry, and Shapiro 2010; Dupas 2014). Ashraf, Berry, and Shapiro (2010) did find that when surveyors offer households free water treatment solution, some households accept it but do not treat their water. However, more recent studies suggest that when the most natural and cost-effective approaches for distributing preventive health products are used, waste is minimal. Pascaline Dupas, Vivian Hoffman, Alix Zwane, and I (2016) find that when households receive vouchers for water treatment solution, the vast majority of those who take the time and effort to redeem the coupons treat their water.

A follow-up study in Malawi found not only that free water treatment vouchers promoted take-up with little waste but found substantial reductions in morbidity over the long run for a very low cost (Dupas *et al.* 2020).

Even when products are free, adoption can still be low, as is the case for childhood immunization in Rajasthan, India. Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Dhruva Kothari (2010) show that a small incentive – distributing lentils during vaccination visits and metal plates for full immunization – substantially improved take-up.

The extensive body of work on preventive health uptake has generated policy insights with widespread implications, and as a result has had a major impact on health policy and outcomes: inexpensive preventive health products such as malaria nets and water treatment solutions are increasingly provided for free, saving lives and improving health for millions around the world. Modeling suggests that mass free distribution of bed nets across subSaharan Africa alone saved 4 million lives between 2000 and 2015 (Bhatt *et al.* 2015). This literature has also raised theoretical questions about whether individuals behave as the fully rational actors of traditional economic models.

Behavioral Development Economics
This pattern of valuable investments left unexploited arise in many sectors other than health. One important example is the low use of fertilizer by many farmers in developing countries.

Esther Duflo, Jonathan Robinson, and I (2011) designed an RCT to test several different theories, both rational and behavioral, of what may be driving the low use. Our findings were most consistent with a model built around an alltoo-familiar form of irrationality: procrastination. This model was based on existing work in behavioral economics (O'Donoghue and Rabin 1999, 2006; Loewenstein, O'Donoghue, and Rabin 2003), and assumed that farmers were not only present-biased but also failed to predict that they would remain present-biased in the future. Because buying fertilizer required time and effort in the present to receive a payoff at a later date, farmers delayed buying until the imaginary future moment when they would finally be patient. The model suggested that offering small, timelimited discounts on fertilizer at harvest time, when farmers had money, would create a "deadline effect" that could help overcome the problem. We tested this and found that such timelimited discounts had as much effect on fertilizer purchases as much larger discounts without the deadline effect.

Work with Kevin Carney, Xinyue Lin, and Gautam Rao (Carney *et al.* 2020) suggests that another nonstandard preference, loss aversion, may interact with naïveté about future preferences in ways that prevent households from taking out loans to invest in capital goods if they have to collateralize those loans with existing assets. In earlier work, William Jack, Joost de Laat, Tavneet Suri, and I (2018) had found that offering dairy farmers the opportunity to collateralize loans to purchase water tanks with the tanks themselves, rather than existing assets, increased take up of the loans from 2.4 percent to 44 percent with very high repayment rates. One hypothesis is that farmers were loss averse and thus unwilling to borrow if it meant putting existing assets at risk, but that since they were loss averse only for goods that they already owned, they were willing to take out a loan to purchase a water tank that would itself serve as collateral. The Jack et al. study was not designed to isolate behavioral factors, but Carney *et al.* (2020) designed a lab-in-the-field experiment with farmers in the area showing substantial loss aversion and naïveté about future loss aversion, and estimated a structural behavioral model indicating that farmers were willing to pay approximately eight percentage points higher interest rates each month to collateralize loans with new assets purchased under the loan, rather than with existing assets.

These examples illustrate the interplay between the experimental approach and behavioral economics. The widespread use of the experimental approach in development economics may help explain why the development field has embraced insights from behavioral economics and the subfield of "behavioral development" (Kremer *et al.* 2019) has flourished. Even when data seem at odds with what basic rational models would predict, sufficiently creative economists can simply add further refinements to these models until observed behaviors could potentially be explained in rational terms. Prior to the widespread adoption of the experimental method, the conversation would often end there. With the advent of the experimental approach, and the development of parsimonious behavioral models, economists could iterate rapidly when they found anomalies. Follow-up experiments could be carefully designed to test specific predictions of various explanations for unexpected findings. As in the case of fertilizer and loan uptake, behavioral explanations often perform better than rational explanations in these follow-up experiments.

Early in his highly influential essay "The Methodology of Positive Economics," Milton Friedman (1953) argues that it is essentially impossible to conduct experiments in economics but that this obstacle only slows progress, rather than meaningfully restricting progress in the long run. He goes on to write that no economic model can – or should – ever incorporate every element of the real world but rather capture its key features[2]. Friedman argues that the ultimate test of an economic hypothesis is its ability to make accurate predictions, and thus realistic assumptions are preferable only insofar as they increase a theory's predictive power. For example, he dismisses the idea of basing economic models of firms on insights gleaned from talking to people in business about how they make decisions, and argues that the assumption that firms are maximizing profits by making decisions based on marginal costs may be best even if businesspeople do not see themselves acting that way.

Friedman then goes on to another step, one that ultimately comes not from his methodological views, but from his beliefs about the world: he argues that models built on the assumptions of rationality and perfect competition do a good job of prediction.

One can only speculate, but it seems likely that Friedman, like some contemporary economists, would be unmoved by laboratory experiments showing departures from rationality. He might see the goal of economics as predicting behavior in real world contexts with parsimonious assumptions, rather than predicting behavior of laboratory subjects playing artificial games with which they have limited experience in environments in which there is no market selection of participants. Field experiments are less subject to this critique.

2. Borges (1998) captures the absurdity of trying to capture all features of the world in his story about an effort to make a perfectly realistic map at a 1:1 scale.

Friedman attributes what he sees as a misguided attachment to realism in part to the difficulties inherent in a science without controlled experiments. Without the ability to design experimental tests of a theory, Friedman contends, it is difficult to convince skeptics, and thus incorrect ideas are abandoned slowly, if at all.

The experimental method has enabled us to quickly test specific predictions coming from behavioral models. It has become clear that models which incorporate assumptions about behavior based on insights from psychology often provide more accurate predictions about real world behavior than more standard economic models. The combination of precise, well-articulated behavioral models and field experiments in development economics has thus allowed for fairly rapid adoption of some behavioral models as fairly standard in the development economics toolkit and to the flourishing of behavioral development economics.

That said, the role of experiments in promoting the rapid adoption of behavioral approaches was not simply because the experimental method allows rapid iteration and causal identification to enable testing of alternative models. The richer sense of context, the focus on specific problems of practical importance, interaction with subjects and with experts in fields outside economics, and the ability to rapidly iterate all played key roles in generating hypotheses and understanding which models to text. In our work on fertilizer, we talked to agricultural scientists, who agreed that most farmers would be better off using at least some fertilizer. We asked farmers about fertilizer. Rather than saying that it was too risky or required too much labor input at times when labor was scarce, they said that they believed fertilizer was profitable, that they planned to use it in the next season, but that they lacked the money to buy fertilizer in the current or previous season.

Historically, many economists would have dismissed the views of both agricultural scientists and farmers, noting that agricultural scientists might be focusing on maximizing yield rather than farmer profits or utility, and that farmers might be angling to get free or subsidized fertilizer. While the experimental method does not imply that one uncritically accepts the views of agricultural scientists and farmers, it makes them harder to ignore, and enables us to test alternative hypotheses informed by their views.

Moreover, because the experimental method gives researchers a better sense of context, it made more salient some facts useful in seeking such a model. For instance, many farmers refer to the period in the run up to harvest as the "hungry season" and prices are highly seasonal rather than being smoothed by grain storage. This is consistent with models of present bias.

The development of the O'Donoghue-Rabin model meant that we could draw on an existing parsimonious model with specific, testable implications. The ability to rapidly iterate meant that we could test those specific predictions against alternative hypotheses, working in a specific context of real-world practical importance. Some see the experimental method as suggesting that evidence and data are more important than models; in contrast, I see the experimental approach as a new tool for generating evidence that can be used to test models, as in the fertilizer paper, or to inspire the creation of new models, focusing on issues we had previously neglected, as in the work on tracking in Kenya.

C. Water, Structural Estimation, and Innovation

The deworming, preventive health, and agricultural investment examples illustrate how the experimental method can be used to shed light not just on specific programs, such as provision of free deworming medicine, but also on more general issues such as underinvestment in cost effective preventive health goods. They also show how experimental research is helping us understand the underlying behavioral factors behind underinvestment more generally. A series of experiments on water illustrate how field experiments can be combined with structural modeling to shed light on questions such as the impact of property rights that may be difficult to address directly with experiments alone; how we can use the experimental method to uncover and at least partially address methodological issues that affect not only experiments but empirical work more broadly; and how an iterative process of experiments, informed by an understanding of behavioral economics, can help produce innovative new solutions to development problems.

Diarrheal disease, often caused by contaminated water, is a major cause of child death in lowincome countries. Ted Miguel, Jessica Leino, Alix Zwane, and I (2011) evaluated an NGO project in Kenya in a region where open springs are an important source of water. These springs are easily contaminated, so the NGO protected them by encasing their sources in concrete.

Both in tradition and in law, people in the area had the right to collect water from springs on their neighbor's land. These types of communal property rights are common around the world. Many economists believe that establishing private property rights can spur investment. So, allowing landowners to charge their neighbors for water might encourage them to invest in upgrades like encasings to make their water sources cleaner.

Under certain assumptions, we could estimate how much people would be willing to pay for cleaner water by measuring how much farther they were willing to walk for it. The experiment allowed us to estimate this parameter, and we found very few people were willing to walk far to obtain cleaner water. This in turn meant land-owners would not be able to charge much for cleaner water, and therefore few would find it profitable to invest in infrastructure improvements. In fact, the parameters from our experiment, combined with our structural model, suggested that creating private property rights for landowners would in this context lead to substantially worse drinking water quality, because people would switch to dirtier public sources, such as streams or lakes, rather than paying for clean water.

This is one example of how field experiments can be combined with other techniques in economics, such as theorybased structural modeling, to address questions such as the impact of different property rights systems, that are themselves less directly amenable to experimentation.

The evaluation of water treatment interventions also led to methodological innovation. Being surveyed more frequently made households more likely to use chlorine at home. There was evidence that in many studies of diarrheal disease, simply being surveyed changed respondents' behavior. In work with more than a dozen other researchers, we found the same effect for many other interventions and other countries: being surveyed changed behavior. This finding has an important practical implication for research and survey design. Where possible, it may be best to use administrative data, rather than surveying program participants. If surveys are needed, we proposed working with larger samples and conducting measurement via surveys less frequently than smaller samples with frequent surveys.

The study also led to policy innovation. We found that spring protection helped: an indicator of fecal contamination at the springs fell by two-thirds and diarrhea rates fell by one-quarter. However, we also found that water often became re-contaminated in storage or in transport to people's homes. This suggested another approach. Treating drinking water with dilute chlorine solution kills microbes, and keeps stored water safe for one to three days. However, though it is effective and available in Kenya and sold via a socialmarketing model, few people buy it.

We tried multiple possible solutions. We tested the social marketing model that was being used to encourage at-home chlorine use at the time, we tried providing coupons for a free trial, we tried providing information. Then drawing on what we had learned, and lessons from behavioral economics, we designed a new solution: a large dispenser of dilute chlorine, put right at the water source. When people collect water, they can add the right dose of chlorine by turning a knob. In designing the dispenser, we drew on our previous studies and from wider lessons from psychology and behavioral economics to try to increase usage rates. The dispenser was salient: big, bright blue, and placed right at the water source. It was incorporated into something people already did – water collection – so use was convenient and could quickly become a habit. It was visible to others, in order to facilitate social norm formation. And importantly, it was provided to customers without charge. The dispenser increased water treatment four-fold. And that increase was sustained when we tested it over the next three years. That innovative approach is now providing clean water for about two million people each day across Kenya, Uganda, and Malawi.

III. INSTITUTIONS FOR INNOVATION

As the examples discussed above illustrate, the experimental approach can be useful both for science and for innovation. By creating an opportunity for more rapid feedback loops, it can allow science and innovation to move forward together, quickly, in a mutually reinforcing process. As we see in biology, and in the market, systems which create variation in approaches and then select for the most successful ones can be a hugely powerful force for change.

Private firms are continuously doing their own experiments in the form of A/B tests, figuring out which innovations work, refining them, and scaling them up. But, there are some areas where the private sector won't invest as enthusiastically, even if innovation could serve important human needs. A key insight from the study of innovation is that social choices around institutions play a key role in determining both the pace of technological change, and whether the direction of this change matches human needs. There is room for innovation in the institutions governing innovation.

Below, I discuss three sets of institutions: those which facilitate experimentation, those through which governments and philanthropists can directly support innovation, and institutions which incentivize the private sector to create specific new technologies, like vaccines, where existing incentives are not fully aligned with social needs.

A. Institutions to Facilitate Experiments

The scale, complexity, and collaborative nature of experimental work means that it requires new institutions. When I first started conducting research in Busia, we had to create everything from scratch. A number of wonderful colleagues and students, including Karen Levy, Carolyne Nekesa, Ted Miguel, and Pascaline Dupas, generously helped build up a local research infrastructure. Our efforts had an important impact in enabling much further research in Kenya. Esther Duflo, Abhijit Banerjee, Iqbal Dhaliwal, and Rachel Glennerster at the Abdul Latif Jameel Poverty Action Lab (J-PAL), and Dean Karlan and Annie Duflo at Innovations for Poverty Action (IPA) systematically created a global research infrastructure. Colleagues in the field have developed many other institutions. Some are based at universities, such as Evidence for Policy Design, Center for International Development (EPoD/CID) at Harvard, the Center for

Effective Global Action (CEGA) at Berkeley, and the I ternational Growth Centre (IGC) at the London School of Economics (LSE) and Oxford. Other independent NGOs have supported scaling of effective solutions, including Give Directly, Evidence Action, Precision Agriculture for Development, and Teaching at the Right Level (TARL) Africa.

The creation of long-run relationships between government departments and researchers holds particular promise. The widespread availability of administrative data, the development of machine learning techniques, and the diffusion of mobile phones, including smart phones, creates opportunities to quickly generate hypotheses about policy solutions, test them, and incorporate the results into policy. Since digital development approaches have very low marginal cost and can be rapidly scaled to reach millions of people, they can deliver very high return on investment even with modest absolute impact. Over time they have the potential to generate even larger benefits as new technologies come online and systems are refined through an ongoing process of iterative experimentation. Digital agriculture is one area where I have been involved, both as a researcher and as a cofounder of an NGO, Precision Agriculture for Development (PAD).

To see this process at work, consider the collaboration between PAD and Ethiopia's Agricultural Transformation Agency. The government had implemented an interactive voice response system that allowed farmers to call in and access pre-recorded agricultural advice messages through menu selection. Though the system reached 2.7 million callers in its first three years, many users hung up before reaching any advice messages. Economists and data scientists from the government and PAD analyzed the data from call logs to understand the bottlenecks in the system and undertook qualitative interviews with farmers, leading to a range of ideas for improvements. These included changing the order of menu options over the agricultural season so that callers were first offered information on issues most likely to be of interest. A/B tests were rapidly conducted, and the most successful ideas were incorporated into the government system. Results were sometimes dramatic. For example, removing the registration required on a farmer's first call increased the share of users accessing content by 21 percent (Walter 2018). Raissa Fabregas, Frank Schilbach, and I (2019) recently undertook a metaanalysis of digital agricultural extension programs, estimating that mobile phone services which provide farmers information on inputs and methods increase yields by approximately 4 percent. While modest in absolute terms, these gains likely generate very high returns for programs that operate at scale, since rural cell phone towers typically operate under capacity, meaning that the social cost of information transmission is close to zero.

Gains are likely to grow over time as smart phones diffuse, lowering costs, enabling delivery of video, and opening up opportunities for provision of information that is more customized, for example by GPS location. Furthermore, because mobile-phone information systems can both collect large amounts of usage data and be quickly and easily modified, they lend themselves particularly well to experimentation. Through experimentation and feedback loops, these systems can be made even more effective over time.

The experience in Ethiopia is typical of PAD's experience in other countries. Governments often design systems with highly technical messages that are difficult for farmers to use or understand but are eager for evidence on ways to improve their systems, and rapidly adopt them (Cole and Sharma 2017). It seems likely that there is space for organizations to work with governments in a range of fields, from agriculture, to education, to health, to tax collection, to the administration of basic utilities and road repair, to find ways to improve services. Such work could potentially be done either by researchers based in universities or in other organizations, but in my view it is important that the results of such work be published, both so it is subject to the quality check provided by peer review, and because the results of evaluations in one country will often be of interest to other countries. Generation and dissemination of this type of information is a global public good.

B. Institutions to Fund Innovation

We need innovation in institutions that fund innovation as well. The World Bank was an early pioneer in this area. Its Strategic Impact Evaluation Fund (SIEF), Development Impact Evaluation (DIME), and Gender Innovation Lab have all been important sponsors of experimental work. The UK's Department for International Development (DFID) supports many research initiatives which fund experimental research, including the International Growth Centre (IGC), Private Enterprise Development in Low-Income Countries (PEDL), and the Agricultural Technology Adoption Initiative (ATAI). The Bill and Melinda Gates Foundation has supported experimental work on microfinance and agriculture. Other funders, such as the Hewlett Foundation, the Wellspring Philanthropic Fund, and the Douglas B. Marshall Jr. Family Foundation have played a key role in supporting the institutions that have made experimental development economics a flourishing field. Other funders, such as Good Ventures, allocate funding in large part based on evidence generated through the experimental method.

More broadly, recent decades have seen a range of initiatives designed to promote innovation in development, including many efforts to invest in natural science research relevant to developing countries, in particular in health, agriculture, and efforts to promote social entrepreneurship and impact investing.

There is a clear theoretical rationale for these initiatives, since innovation is a global public good, and hence is likely to be undersupplied by both the market and national governments. However, there are certainly examples of both success and failure, and one critical empirical question is whether these initiatives are a good use of development funding relative to other investments.

While there is a tradition of assessing the return on scientific R&D dating back to Griliches (1958), the evidence on development innovation investments is largely limited to anecdotes and simple counts of successes versus failures. It is typically difficult to assess whether the social return on a development innovation portfolio exceeds those on alternative investments because it takes years for innovations to scale, data on innovation scale and impact are often unavailable, and many innovations generate social benefits that are difficult to value in dollar terms.

Sasha Gallant, Olga Rostapshova, Milan Thomas, and I (2019) developed an approach to addressing this question and applied it to data from the early years of the United States Agency for International Development's Development Innovation Ventures (DIV)[3]. The approach is based on the idea that, under fairly plausible assumptions, even in the absence of data on the rate of return on many investments, one can obtain a lower bound on the rate of return on a portfolio by comparing the benefits of a subset of investments to the cost of the entire portfolio.

DIV supports innovations for development from a wide variety of sources, including social entrepreneurs, for-profit firms, and researchers. Unlike most innovation funders, DIV is deliberately open across sectors and geographies, and to innovations intended to scale either commercially or through developing country governments or donors. This openness is complemented with a tiered, evidence-based approach to funding. DIV makes small investments to pilot and test promising ideas, and larger ones to help transition to scale innovations that are supported by rigorous empirical evidence of impact and cost effectiveness or that have demonstrated market viability. Like many research funders, but unlike many impact investors, DIV's decision-making process involves peer review, often by development economics researchers.

Because innovations take time to scale, we focused on DIV's early portfolio: the 41 innovations that received grants between 2010 and 2012. In general, investments in research and innovation yield highly skewed outcomes, and the evidence from DIV is consistent with this pattern.

3. Disclosure: I cofounded and serve as Scientific Director for DIV.

Over 95 percent of the 56 million direct users reached by innovations in DIV's early portfolio through the end of 2018 were due to the top nine innovations, with the top quartile of the innovation reach distribution approximated by a power law distribution.

This skewness means that a bounding approach can potentially be informative. Of the nine innovations reaching more than one million users, we were able to obtain plausible estimates on the social benefits generated by four. Through 2018, the discounted social benefits of these four early portfolio innovations were more than five times as large as the discounted cost of the entire portfolio (award values and estimated administration). The actual benefit/cost ratio is likely much greater, since the benefits of most supported innovations were not included, and innovations will continue to deliver benefits after 2018. This estimate provides evidence that the open, tiered, and evidence-based approach to investing can yield large social returns.

An analysis of which innovations in the early DIV portfolio reached over one million users – though based on a small sample – is illustrative of which types of innovations are likely to scale in general[4]. While early-stage pilots were less likely to have scaled than later stage innovations, they also cost much less, so the point estimate of the number of users reached per dollar spent was actually greater for early stage innovations, contrary to the aphorism in development that "pilots never scale."

Cost was a key predictor of scaling: innovations were three times more likely to scale if they had estimated unit costs less than US$3 per person. Perhaps because of this, innovations were three times more likely to scale if, rather than trying to sell directly to consumers, they leveraged pre-existing distribution networks (for example by working with a government or existing large-scale business). Customer acquisition costs are a major obstacle when trying to sell items that cost less than US$3.

The analysis of predictors of scaling also underlined the role of the experimental method in promoting innovation. Projects that had experimental evidence of impact and cost effectiveness at the time of application to DIV were five times more likely to scale than those that did not. Projects involving development economics researchers were also more likely to scale than those that did not.

There are many potential explanations for this correlation between research using the experimental method and scaling. Rigorous experimental evidence may spur adoption by institutions. Perhaps innovators only undertake external trials when they are confident of success, and the correlation reflects a selection effect. But the data are also consistent with the hypothesis that researchers using the experimental method can be

4. This followed an earlier analysis by Duflo and Kremer (2015).

successful innovators. For example, James Habyarimana and Billy Jack developed an approach to reducing the horrific rate of accidents in minibuses in Kenya by posting stickers inside buses. The stickers encourage passengers to speak up against reckless driving. It was piloted and tested with the support of DIV, and a randomized controlled trial showed that the stickers reduced insurance claims by 25 percent (Habyarimana and Jack 2015). In Kenya, the innovation was scaled up by an insurance company (which requires stickers as a condition for coverage), and the government. To take another example, Asim Khwaja and Bailey Klinger developed a mobile phone-based psychometric credit scoring program to assess default risk, enabling banks to lend money to people who may not have had credit history and collateral (Klinger, Khwaja, and del Carpio 2013). Banks used the approach to make over $1.5 billion in loans to over one million customers. Researchers in IT and biomedicine often develop innovations and sometimes help take these to scale, and this is now happening in development economics as well.

Taken together, the magnitude of the social return on the early portfolio and correlates of innovation scale suggest that there are arbitrage opportunities for socially motivated funders to invest in innovation in areas that would otherwise be neglected by the private sector. With reasonably well-functioning markets, the private sector will claim most innovation investment opportunities that are both profitable and socially beneficial (e.g., mobile phones). That leaves opportunities for socially motivated innovation funds to deliver large social returns by identifying opportunities for which the ratio of expected social to private return is high.

Standard industrial organization theory suggests that private returns to innovation may be a higher fraction of social returns in settings where barriers to entry are high. This implies that there may be arbitrage opportunities for investors motivated by social returns in areas where there are few barriers to entry, such as early-stage innovation, or innovation that is being conducted by researchers whose results will be put into the public domain. Private investors may have difficulties fully appropriating the returns from innovations that are only valuable in combination with a proprietary asset owned by another entity. Therefore, there may be arbitrage opportunities for investors seeking social returns to invest in innovations intended to scale through adoption by governments or large firms that would have substantial market power in bargaining with the innovator.

C. Alternative Ways of Rewarding Innovation

Upfront "push" financing of innovation will always potentially be subject to asymmetric information problems and potentially to political economy problems. Governments therefore complement push funding systems with "pull" institutions designed to reward successful innovations. In particular, patents and other intellectual property right systems reward innovators with monopolies over the use of their ideas. Monopoly pricing creates the usual static distortions. Moreover, even with monopoly power, dynamic distortions remain. Monopolists who cannot pricediscriminate will be unable to fully capture consumer surplus or knowledge spillovers generated by research. Chris Snyder and I (2015) show that if demand curves have a truncated Zipf-like shape, monopolists may be able to extract an arbitrarily small share of social value and argue that demand curves for preventives are often fairly Zipf-like, reducing incentives for innovation in preventives relative to treatments. Patents also encourage socially wasteful expenditures to create "me too" products: slight variations on existing products which differ just enough to avoid patent coverage.

In "Patent Buyouts: A Mechanism for Encouraging Innovation" (1998), I describe an auction system which governments could use to determine the market value of patents. With low probability the auction would be implemented, but with higher probability the government would offer to purchase the patent from the innovator at some fixed markup over the high bid in the auction and put the patent in the public domain. Such a system would increase incentives for original innovation, help eliminate "me too" research, and prevent monopoly pricing. Of course, it might be subject to a host of practical problems.

For some goods, such as vaccines, it is relatively easy to determine the socially efficient quantities and even to assess the value of products, using the randomized trials required by the regulatory system. In Strong Medicine: Creating Incentives for Pharmaceutical Research on Neglected Diseases (Kremer and Glennerster 2004) we proposed another pull funding instrument, Advance Market Commitments (AMC). The proposals in Strong Medicine were focused on vaccines in developing countries. Under an AMC, donors commit in advance to "top-up" payments for a certain number of vaccines by a pre-agreed amount, if a firm develops a vaccine meeting technical specifications, and agrees to set the long-run price close to marginal cost.

AMCs aim to address several market distortions. Top-up payments bolster dynamic R&D incentives by guaranteeing firms a certain price per dose, increasing the likelihood that a vaccine is developed. The long-run price cap addresses static distortions from monopoly pricing, effectively enabling two-part pricing.

In 2007, a consortium of donors pledged $1.5 billion to set up a pilot AMC for a pneumococcal vaccine covering the strain of the disease common in developing countries. At the time, pneumococcus killed 1.6 million people each year, primarily in developing countries. Three firms have now developed pneumococcal vaccines covering the strains of the disease common in developing countries. Vaccines have now been distributed to more than 150 million children, saving an estimated 700,000 lives.

Kremer *et al.* (2020) discusses the experience of the pneumococcus AMC pilot, noting that firms may invest in less than the socially optimal amount of capacity in the absence of an AMC, and arguing that the details of the design of an AMC will be critical in ensuring that they indeed incentivize capacity creation. We show that the rate of pneumococcus vaccine coverage in eligible countries converged to the global rate almost five years faster than for rotavirus vaccine, which was not supported by an AMC. We argue that based on the results of the pneumococcus AMC, it would be worth experimenting with AMCs for other technological targets – including more technologically distant targets – and discuss possible refinements that could be used to improve the functioning of AMCs.

IV. CONCLUSION

I have argued that the experimental approach can be a powerful new tool for advancing economic science, for informing policymakers, and for stimulating innovation. This is both because experiments enable more credible isolation of causal impact from potential confounding factors, and because experiments give the researcher a richer sense of context; address specific, practical problems; promote broader collaboration, and allow more rapid iteration.

Experiments have proven to be applicable to a much broader range of problems than I initially imagined when I started working on education in Busia, Kenya. These include questions of political economy (Callen and Long 2015, Wantchekon 2003); corruption (Olken 2007); democratic institutions (Bidwell, Casey, and Glennerster forthcoming); censorship (Chen and Yang 2019); learning by exporting (Atkin, Khandelwal, and Osman 2017); women's empowerment (Chattopadhyay and Duflo 2004); and labor markets (Breza, Kaur, and Krishnaswamy 2019) to name just a few. The techniques have been enriched by combining field experiments with lab experiments (Ashraf, Karlan, and Yin 2006; Kaur, Kremer, and Mullainathan 2015) and structural modeling (Todd and Wolpin 2006, 2010; Attanasio, Meghir, and Santiago 2012; Duflo et al. 2018; Kaur, Kremer, and Mullainathan 2015). Methodological challenges have been identified and techniques have been developed for at least partially addressing them (Casey et al. 2012, Rosenzweig and Udry 2020).

But this is just the beginning. We need to encourage a variety of new approaches to take advantage of new opportunities, test new approaches, refine them, and scale up the most effective solutions.

As Esther, Abhijit, and I have emphasized, this is a prize not just for the three of us but for a broader movement involving researchers, nonprofit organizations, businesses, and governments that want to improve what they are doing for the broader good and are courageous enough to innovate and rigorously test their innovations. This movement has much left to accomplish, and plenty of room to grow. It requires many people in many different roles, be they researchers or survey enumerators, entrepreneurs, or policymakers. The experimental approach is itself an innovation which drew on earlier economic research and on work from other disciplines. Through the combined efforts of those using the experimental approach and those using other methods, we understand issues in development economics much better today than we did when I began my career. I look forward to the new methodological innovations which will build on the insights of the experimental approach, further deepen our understanding of development, and help improve lives around the world.

Acknowledgments

1. I am very grateful to Amrita Ahuja, Arthur Baker, Oriana Bandiera, Owen Barder, Jeff Brown, Robin Burgess, Shawn Cole, Jonathan Faull, Rachel Glennerster, Matthew Goodkin-Gold, Tomoko Harigaya, Asim Khwaja, Dan Levy, Fanele Mashwama, Stephen Nyarko, Dev Patel, Mala Persaud, Marla Romash, Milan Thomas, and Junyi Que, for their extremely helpful comments and suggestions. I would like to thank my coauthors for all that they have taught me over the years: Daron Acemoglu, Amrita Ahuja, Lorena Alcázar, Joshua Angrist, Sarah Baird, Abhijit Banerjee, Owen Barder, Ernst Berndt, Eric Bettinger, Olivier Blanchard, Hoyt Bleakley, Erik Bloom, Johanne Boisjoly, Adrien Bouguen, Conner Brannen, Ben Broadbent, Ryan Bubb, Donald Bundy, Lorenzo Casaburi, Marcos Chamon, Nazmul Chaudhury, Daniel Chen, Michael Clemens, David Clingingsmith, Jessica Cohen, Bruno Crépon, Kevin Croke, David Cutler, Joost de Laat, Florencia Devoto, Esther Duflo, Greg Duncan, Pascaline Dupas, William Easterly, Jacque Eccles, Raissa Fabregas, Willa Friedman, Winnie Fung, Sasha Gallant, Xavier Ginéj, Rachel Glennerster, Paul Glewwe, Mary Kay Gugerty, Jeffrey Hammer, Joan Hicks, Vivian Hoffmann, Alaka Holla, Richard Hornbeck, Hsu Hsu, Yue Huang, Nauman Ilias, William Jack, Seema Jayachandran, Matthew Jukes, Dean Karlan, Supreet Kaur, Asim Ijaz Khwaja, Elizabeth King, Jean Lee, Jessica Leino, Jonathan Levin, David Levine, Ruth Levine, Dan Levy, Isaac Mbiti, Atif Mian, Edward Miguel, Ashoka Mody, Charles Morcom, Sylvie Moulin, Sendhil Mullainathan, Karthik Muralidharan, Robert Namunyu, Clair Null, Benjamin Olken, Alexei Onatski, William Pariente, Catherine Patillo, Shawn Powers, Lant Pritchett, Gautam Rao, Jonathan Robinson, Halsey Rogers, Olga Rostapshova, Juan Saavedra, Frank Schilbach, Monica Singhal, Christopher Snyder, James Stock, Lawrence Summers, Tavneet Suri, Milan Thomas, Rebecca Thornton, Eric Van Dusen, Tom Vogl, Heidi Williams, Jack Willis, Eric Zitzewitz, Alix Zwane. I am also very grateful to my Kenya-based collaborators, including Carolyne Nekesa, Karen Levy, Suleiman Asman, and Chip Bury. I would also like to thank all of my students. For their support and inspiration, I would also like to thank my wife Rachel Glennerster, my sons Ben and Daniel, my parents, Sarah Lilian Kremer and Eugene Kremer, my advisors Robert Barro and Peter Timmer, and my childhood mentors John Eck and Larry Weaver.